International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Securing Autonomous AI Agents in Industry 5.0: A Review of Analytics-Driven Identity and Access Control

Raj Savani*, Krina Mashru**
*(Computer Engineering, Atmiya University, Rajkot, Gujrat, India
Email: therajsavani@gmail.com)
*((Faculty of Engineering and Technology (CE), Atmiya University, Rajkot, India
Email: krina.masharu@atmiyauni.ac.in)

----------**********----------

Abstract

The implementation of autonomous AI agents in Industry 5.0 cybers-physical systems offers a new layer of security concerns, in which the traditional and static identity and access management (IAM) model would be insufficient. This literature review summarizes the existing study on the potential risks posed by AI agents that include policy manipulation, unauthorized agency, and data poisoning and assesses the development of analytics-driven IAM and Zero Trust models aimed to address them. We discuss adaptive controls to continuously verify (with risk) non-human identities, on the experience of cyber-physical deployments. Other sophisticated monitoring methods including anomaly detection through inverse reinforcement learning are also reviewed in order to assure agent behaviour. Lastly, we determine that the standardized datasets, observability pipelines, and formal verification processes have critical gaps, which will result in the future research agenda to ensure the safe and reliable use of AI agents in industries.

Keywords — **AI Agents, Identity Management, Access Control, Cybersecurity, Industry 5.0, Zero Trust.**

----------**********----------

I. INTRODUCTION

The introduction of autonomous AI agents into Industry 5.0 cyber-physical systems is the paradigm shift between automated and cognitive manufacturing and allows the collaboration of humans and machines never before seen [1], [2]. These can be seen as agents that are both changing the way industrial processes are run but are also posing new security challenges that conventional models of security fail to handle effectively [3].

Contrary to traditional software, AI agents work as the dynamic non-human identities whose behaviours and policies change over time [4]. This inherent property makes the role-based access control systems that are static systems outdated, and it introduces weaknesses were adversarial manipulation, abuse of policy, or other agencies without permission can have serious safety or security risks in industrial contexts [5], [6].

The cybersecurity community in their turn is moving towards analytics-based Identity and Access Management (IAM) and Zero Trust Architecture that offers sustained, context sensitive security controls [7], [8]. These structures move to the model of single authentication to that of continuous verification on behavioural analytics and risk assessment in real time. Other security layers provided by complementary monitoring methods such as anomaly detection and assurance frameworks are added to autonomous systems [9], [10].

The paper will give a detailed literature review of the literature on ensuring the security of AI agents in Industry 5.0 settings. We consolidate the research on the distinct threat environment, examine the new analytics-based IAM and Zero Trust solutions, and survey monitoring and assurance methods. The review ends with the definition of crucial research gaps and a specific agenda that will allow the safe implementation of autonomous AI agents in industrial ecosystems.

The paper is organized in a way that it will first analyze the threat environment (Section III), give integrated security frameworks (Section IV), and lastly discuss monitoring ecosystems (Section V) and research gaps and conclusions will follow.

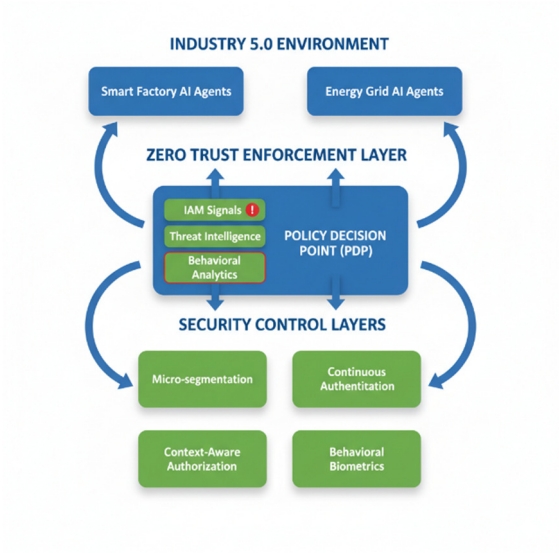

Fig. Integrated AI Agent Security Framework for Industry 5.0, showing the layered defense architecture combining analytics-driven IAM, Zero Trust enforcement, and security controls.

II. REVIEW METHODOLOGY

The review used a systematic method to find and analyse literature at the cross-point of AI agent security, identity management, and Industry 5.0. The methodology was effective because it was able to cover a wide range of the research but at the same time stayed on the most pertinent and effective research.

The literature search has been made in key academic databases such as IEEE Xplore, ACM Digital Library, and arXiv in order to cover both peer-reviewed articles and powerful pre-prints. Search terms: Key terms were used together, i.e., AI agent security, non-human identity, autonomous system access control, Zero Trust AI and Industry 5.0 cybersecurity.

The selection of papers was done using the following criteria:

- **Inclusion:** 2022-2025 publications that talk of the security, identity or access control of autonomous AI agents especially when they are talking about analytical or adaptive security models.
- **Exclusion:** Articles that discuss the traditional IT security solely, without autonomous agents, or articles that are not published in English.

The result of this procedure was a fundamental set of 11 scholarly articles [1]-[11] that establish the basis of the present review. To align the academic literature with the industrial practice, three high impact industry and government reports [12]-[14] were included, which were chosen due to their relevance to the operational AI agent deployment issues and publishing by recognized authorities. The following paragraphs give thematic synthesis grouped around three main areas namely the changing nature of threats (Section III), analytics-based security models (Section IV) and monitoring/assurance ecosystems (Section V).

III. THREAT LANDSCAPE OF AI AGENTS

Autonomous and adaptive AI agents pose a different security risk, which goes beyond the traditional software threats. These hazards attack the fundamental paradigms of operations of agents their decision-making policies, training data and autonomy. On our analysis, we classify these threats into four major categories as summarized in Table I.

TABLE I
TAXONOMY OF AI AGENT-SPECIFIC THREATS

Threat Category	Attack Vectors	Industry 5.0 Impact
Policy Manipulation & Adversarial Attacks	Adversarial inputs, reward poisoning [8], [9]	Physical damage, production defects
Data Integrity & Model Poisoning	Training data injection [2], [3]	Performance degradation, unsafe actions
Unauthorized Agency & Privilege Escalation	System integration exploits, prompt injection [13]	Data exfiltration, unauthorized access
Provenance & Impersonation Risks	Identity spoofing, audit trail gaps [7], [14]	Accountability loss, cascading failures

Policy manipulation is a serious menace in which enemies take advantage of the learning processes of the agent. Subtle manipulations of sensory inputs can cause disastrous decisions in the agents of the reinforcement learning, as has been shown by Lian et al. [8] and Muller et al. [9]. In the industrial environment, this may be in the form of tweaked visual cues that make autonomous

vehicles run in the wrong path or sensor measurements that are distorted and confuse quality-control measures.

The agents are compromised by the data integrity attacks during the training stages to form Trojan horse models that act normally until aroused. Hendrycks et al. [2] list this as a catastrophic AI risk, and Raza et al. [3] observe that multi-agent systems have the potential to increase the effects of poisoning in whole ecosystems. A manufacturing agent that has learned about poisoned data may slowly learn to maximize dangerous production speeds.

Unauthorized agency is a result of the basic essence of AI agents as autonomous non-human identities [7]. The DTEX threat advisory [13] outlines the so-called Lethal Trifecta that allows prompt injection attacks to override original instructions, which makes supply chain managers the means of shipment diversion or data exfiltration.

Provenance risks are a problem of accountability in multi-agent setting. Chen and Chen [7] emphasize the challenges in understanding the legitimate and malicious bots, whereas the Australian policy [14] focuses on the necessity of clear AI systems. In the absence of a strong attestation, blackouts may be caused by a malicious agent who can assume the role of a grid controller.

IV. ANALYTICS-DRIVEN IAM AND ZERO TRUST FRAMEWORK

The dynamism of the AI agents demands the shift of the traditional security models to dynamic and continuous models. This evolution centers on two orthogonal directions: Identity and Access Management (IAM) that is analytics-driven and dynamically-oriented to be able to validate the identity and Zero Trust Architecture (ZTA) to implement the policy in a granular fashion.

The frailty of traditional IAM is gauged by the reality that based on the industry statistics 96 percent of organizations have complications with non-human identities [12]. This identity automation divide is the driving force of three fundamental innovations:

1. ***Continuous Authentication***
   It does not require any fixed credentials but rather the live trust assessment, which authenticates the digital signatures, container integrity and network patterns throughout all the agent lifecycle phases [6], [13].
2. ***Behavioural Biometrics***
   This capitalizes on the repetitiveness of the behaviour of the agents e.g. sequence of API calls, decision latency, resource consumption, to generate unique behavioural fingerprints. Research by Lian et al. [8] and Muller et al. [9] suggests that detection of anomalies could be reapplied in the security enforcement system where automatic re-authentication was achieved in the event of behaviour deviation.
3. ***Context-Aware Authorization***
   It will also adopt real time risk scoring that will be founded on the situations of the operation- time, location, sensitivity of the action requested and access decision will be adaptive and not binary one will be [6], [14].

Zero Trust Architecture provides means of enforcement of these dynamic policies in two significant aspects:

1. ***Micro-segmentation***
   It creates closed environments of agent communities, including the possible breaches and blocking the horizontal flows [6], [13]. The control system of an industrial robot is still logically divided into control and financial analysis agents and all inter-zone communication is explicitly approved.
2. ***Dynamic Policy Decision Points (PDPs)***
   It combines IAM signals with threat intelligence signals as well as behavioural analytics signals to support real-time, risk-adaptive access decisions [6], [12]. A PDP may demote agent privileges when they identify abnormal behaviour patterns or time of access.

The collaboration between IAM enabled by analytics and Zero Trust builds a security fabric which adapts alongside the agents that it secures and shifts the assumptions of perimeter to evidence-based management of trust.

V. MONITORING AND ASSURANCE ECOSYSTEM

The arms race that AI defenses are adapting demands that we bring in a new paradigm wherein we apply AI to protect AI against AI, in other words, intelligent monitoring systems identify and respond to AI controlled threats on autonomous environments. It is a system wide immune response more of an ecosystem in that disparate signals are correlated to identify advanced attack patterns that circumvent preventative actions.

This approach is based on algorithmic monitoring. As illustrated by Lian et al. [8], Inverse Reinforcement Learning (IRL) techniques can be applied to make sure that security systems learn the intent of the agent and identify behavioural deviation that indicates that the policy is being manipulated. Related to that, Müller et al.

[9] show that the loop of the reinforcement learning process can also exhibit abnormalities, which can be used to signify the presence of training data poisoning or adversarial interference in the process.

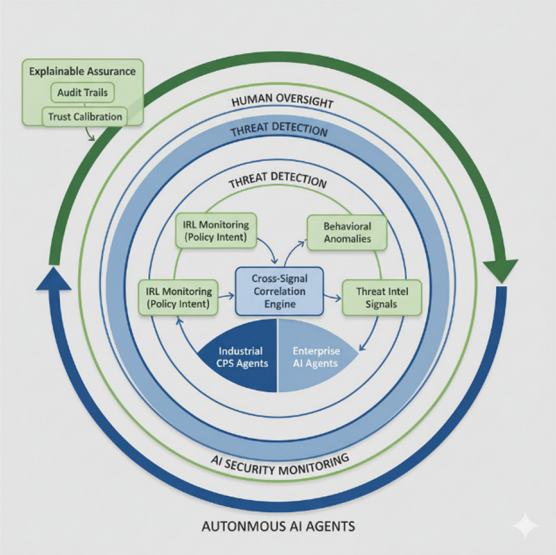

Fig. 2 AI-to-AI security monitoring ecosystem. The diagram shown above depicts the ongoing process of detecting threats, correlating cross-signals, and human control over autonomous agents to ensure their safety.

When such individual signals coincide, the actual defensive force is obtained. An amalgamation of behavioural abnormalities [8], identity context [7] identity and threat intelligence [13] results in an evaluation risk, which is multi-dimensional and enables autonomous response. On detection of suspicious behavioural changes by an IRL monitor and the identity systems detecting abnormal access pattern, the Policy Decision Point can automatically isolate or happily sandbox the possibly breached agent, before it can fall prey to a full-scale breach.

This kind of technical surveillance must ultimately have a positive impact on the spirit and perpetuity of the operations as far as human trust is concerned. The TRiSM framework [3] emphasizes the reality that security controls are not supposed to compromise human confidence through comprehensible actions. Predictability and transparency form the basis on which trusting is established as presented by Vanneste and Puranam [4]. Therefore, since this system will be independent in handling a threat, it must be capable of providing a transparent audit trail with behavioural scores, contextual violations, and identity information to clarify to human operators.

The full security cycle is then closed: AI-generated threats are identified by AI systems with the help of AI-driven analytics, and human control is ensured by AI-explainable assurance. This produces a robust, dynamic security position with defence developing at the same rate as the threats against it.

VI. RESEARCH GAPS AND FUTURE DIRECTIONS

The extrapolation of the current literature demonstrates that the success of the AI agents will not be the result of the mere improvement, but the innovative creation that will allow bridging three major gaps hindering the process of transforming the theoretical framework to the real world.

The Benchmarking persists in the fact that there are no uniform datasets and measures of AI agent security. Despite the potential of such methods as IRL-based monitoring [8] and behavioural biometrics [9], there are no standards within the area which allow to test the level of real-world implementation of them. This is where the problem originates that is identified by the identity automation gap described by SailPoint [12] as it cannot be guaranteed to have autonomous systems without the proven tools which can be measured.

The Observability Gap does not allow the achievement of the integrated security ecosystem in Section V. Data silos between identity systems, behavioural monitors, network sensors, and threat intelligence form blind spots that can be used by advanced attackers. As Talakola [5] observes in the broader analytics settings, the prerequisites of actionable insights are unified data pipelines but no standards are defined to monitor agent behaviour or cross-signals correlations.

The Assurance Gap symbolizes the unbridgeable gap between detectable safety and verifiable safety. Today, monitoring can identify abnormalities [8], [9] and is not capable of formal guarantees and diagnosis in an explainable fashion. This is also reflected directly on human trust [4] and not within the regulatory standards of transparent AI [14]. TRiSM framework [3] is not quantified nor does it have automated enforcement systems, it is merely a conceptual framework.

In order to fill in these gaps, we offer three interrelated research projects:

1. ***AgentSec-Bench***

   A community-based benchmarking set of taxonomies of standardized attacks, performance metrics and representative scenarios of Industry 5.0.

2. ***Unified Observability Platforms***

   Open-source pipelines and data models that normalize telemetry across identity, behaviour, and threat dimensions, enabling the

cross-signal analysis essential for detecting multi-stage attacks.

3. ***Explainable Assurance Systems***

Ways to transform detection events into human interpretable diagnoses and provide lightweight formal verification of agent safety properties, implementing TRiSM through standardized risk API.

This agenda recognizes that autonomous AI must be achieved via security being not a feature but an emergent feature of intelligently designed ecosystems in which prevention, detection and assurance constitute an ongoing process of adaptive cycles.

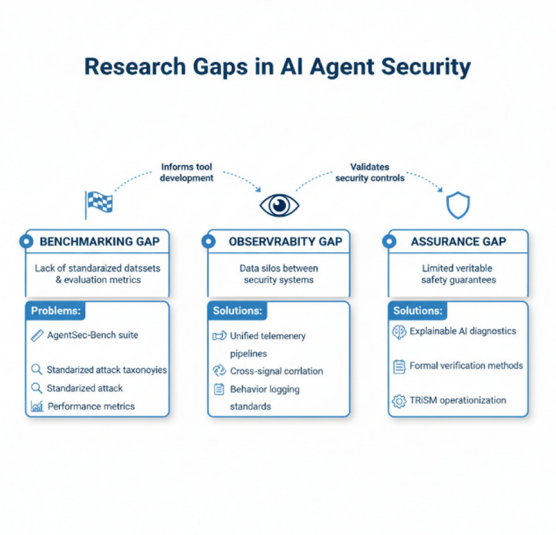

Fig. 3 Gaps in critical research in the field of AI agent security: benchmarking, observability, and assurance issues with solution initiatives.

VII. CONCLUSIONS

This review has been critical in discussing the paradigm shift that needs to be done to ensure that autonomous AI agents are present in Industry 5.0 situations. In the context of the analysis, it is possible to note that the conventional approaches to security, which rely on the principle of pre-defined permissions and anthropomorphic identities, cannot be discussed as the functioning ones in the framework of dynamic and adaptive AI systems. The evidence-based trust management is a novel philosophy of security, which has to be perpetually evidence-based, as opposed to the so-called assumed perimeter-based security, in excessive numbers, as the testaments expression goes.

The literature tour has revealed that such AI agent threats like policy manipulation and data poisoning, unauthorized agency, etc. are issues that demand a set of defence mechanisms. Dynamic identity verification is provided by the foundation of IAM based on analytics and granular and contextual enforcement is achievable with the foundation of zero trust architecture. Above all, enhanced surveillance implies AI-assuring-AI universe, where behavioural analytics, anomaly recognition and assurance models are a single security meshwork.

Nevertheless, to fulfil this vision, the issues behind benchmarking, observability and verifiable assurance should be considered. The proposed research will have the standardized evaluation as part of the targeted research agenda that will give a clear direction of what to do based on the propose research. It becomes possible to change the direction towards responding to threats and actively creating resilient environments because security is not seen as an added value that can be added to systems, but as the qualitative feature of systems that have been designed intelligently.

Finally, the safe implementation of AI agents in Industry 5.0 will have as a foundation our capacity to develop security ecosystems with the possibility of being able to keep up with the pace of change of the autonomous system, on which they will be implemented. It does not only need technical invention, but also a total re-definition of the principle of trust, identity and accountability in the conditions of human-AI work. The only question that has been raised to the research fraternity now is how these conceptualized frameworks find their way into operational reality.

ACKNOWLEDGMENT

The authors owe their debt to the research community in the field of AI security and Industry 5.0 whose work enabled them to conduct this extensive review. We also value the information provided by identity management experts and zero trust architectures that were valued in our analysis.

The author also thanks Prof. Krina Mashru, Faculty in Computer Engineering Department of Bachelor's, Atmiya University, for their constant support, motivation and guidance in this whole process.

REFERENCES

[1] He, Y., Wang, E., Rong, Y., Cheng, Z., & Chen, H. (2025). Security of AI agents. In 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE) (pp. 45-52). IEEE.

[2] Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv preprint arXiv:2306.12001.

[3] Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025). TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. arXiv preprint arXiv:2506.04133.
[4] Vanneste, B. S., & Puranam, P. (2024). Artificial intelligence, trust, and perceptions of agency. Academy of Management Review.
[5] Talakola, S. (2022). Analytics and reporting with Google Cloud platform and Microsoft Power BI. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(2), 43-52.
[6] Paul, E. M., Mmaduekwe, U., Kessie, J. D., & Dolapo, M. (2024). Zero trust architecture and AI: A synergistic approach to next-generation cybersecurity frameworks. International Journal of Science and Research Archive, 13(2), 4159-4169.
[7] Chen, Z., & Chen, D. (2023). Identifying the Invisible: A Comprehensive Approach to Distinguishing Software Bots.
[8] Lian, B., Kartal, Y., Lewis, F. L., Mikulski, D. G., Hudas, G. R., Wan, Y., & Davoudi, A. (2022). Anomaly detection and correction of optimizing autonomous systems with inverse reinforcement learning. IEEE Transactions on Cybernetics, 53(7), 4555-4566.
[9] Müller, R., Illium, S., Phan, T., Haider, T., & Linnhoff-Popien, C. (2022). Towards anomaly detection in reinforcement learning. In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (pp. 1799-1803).
[10] Alowaidi, M., Sharma, S. K., AlEnizi, A., & Bhardwaj, S. (2023). Integrating artificial intelligence in cybersecurity for cyber-physical systems. Electronic Research Archive, 31(4).
[11] Nweke, L. O., & Yayilgan, S. Y. (2024). Opportunities and challenges of using artificial intelligence in securing cyber-physical systems. In Artificial Intelligence for Security: Enhancing Protection in a Changing World (pp. 131-164).
[12] SailPoint. (2024). The Identity Automation Gap: 96% of Organizations Struggle to Secure Non-Human Identities. [Industry Report]
[13] DTEX Systems. (2024). The Lethal Trifecta: Advanced Threats Targeting AI Assistants and Agents. [Threat Advisory]
[14] Australian Government, Department of Industry, Science and Resources. (2024). Safe and Responsible AI in Australia. [Government Policy]

