

Designing A Least-Privilege Capability Framework for Tool-Using AI Agents in Enterprise Software Systems: Securing Autonomous Operations through Policy-Based Authorization

Durga Narayana Varma Addepalli

Submitted: 07/03/2024 Revised: 19/04/2024 Accepted: 25/04/2025

Abstract: The adoption of tool-using AI agents into the enterprise environment rapidly presents highly influential security issues because of the abundance of permissions and unregulated access to resources. The study suggests a Least-Privilege Capability Framework on AI Agents (LPCF-AI) in which the LLM reasoning and deterministic authorization control are decoupled and task-specific access management is achieved. Design Science Research design using machine learning in the form of risk assessment theories of Logistic Regression, Random Forest, and Gradient Boosting is used. Experimental findings indicate that LPCF-AI decreases the permissions by 76% and still is effective in its work, where the Logistic Regression has the accuracy of 0.691 and the ROC-AUC of 0.749. The framework enhances security, transparency, and governance of autonomous AI activities.

Keywords: *Agentic AI, Least-Privilege Access Control, AI Security, Capability-Based Authorization, Enterprise AI Governance, Zero Trust, Policy-Based Authorization, Machine Learning Risk Prediction.*

I. INTRODUCTION

Background and Context

The rapid development of Large Language Models (LLM) has led to the rapid adoption of autonomous AI agents, which can fulfill complex tasks of an enterprise. The use of AI agents that use tools is starting to be utilized in software development, cloud computing, database administration, and cyber-security operations [1]. As the systems are increasingly equipped to make more decisions and to operate autonomously, the topic of secure interaction with the resources of the enterprise has become a significant research issue.

Problem Statement

Contemporary AI agents must have access to various resources in an enterprise, such as API, database, repository, and cloud infrastructures to perform autonomous work. Nevertheless, the current Identity and Access Management (IAM) frameworks mainly provide services to human users and non-adaptive

applications as opposed to dynamic AI systems [2]. Excessive permissions result in threats including information leakage, privilege growth, unauthorized amendments and supply-chain attacks that pose a great security risk.

Research Aim

The study aims to create a least-privilege capability architecture that enables AI agents to safely operate in an enterprise environment by dynamically granting minimal necessary permissions without impacting their functionality.

Research Objectives

- *To explore issues of insecurity with regard to tool-using AI agents.*
- *To explore the current access-control methods to autonomous systems.*
- *To design a least-privilege framework of AI agents on a capability's basis.*
- *To assess the effectiveness of the policy-based authorization in mitigating security threats.*

Independent Researcher, USA

Novel Contribution

This study presents a least-privilege capability model of tool-using AI agents disentangling LLM reasoning with deterministic enforcement of authorization. In contrast to the view of traditional access-control models, the offered approach allows allocating

permissions based on the tasks regarding the enterprise resources, such as APIs, databases, cloud platforms, and repositories [3]. The framework improves the security of AI agents by limiting the unnecessary privileges and keeping operation in terms of autonomy, transparency, and governance.

TABLE 1. COMPARISON WITH EXISTING ACCESS-CONTROL APPROACHES

Approach	Permission Strategy	Limitation for AI Agents	LPCF-AI Improvement
RBAC	Permissions assigned according to predefined roles	Static roles may provide unnecessary privileges for dynamic AI tasks	Generates task-specific capabilities based on required actions
ABAC	Access decisions based on attributes and contextual information	Requires predefined attributes and policies; difficult for autonomous tasks	Dynamically creates minimum capabilities according to task objectives
Policy-Based Authorization	Uses security policies for access decisions	Policy enforcement is separated from adaptive AI reasoning	Combines policy validation with capability generation and continuous risk assessment
LPCF-AI	Task-based capability authorization	Designed specifically for tool-using AI agents	Separates LLM reasoning from deterministic authorization enforcement

Unlike the traditional IAM principles, LPCF-AI doesn't enable AI agents to directly control enterprise resources. In contrast, the LLM carries out reasoning and task planning-and a dedicated authorization layer verifies generated capabilities prior to their execution. This separation is to minimize risks from autonomous reasoning errors and over permissions.

II. RELATED WORK

Security Challenges and Vulnerabilities in Tool-Using AI Agents

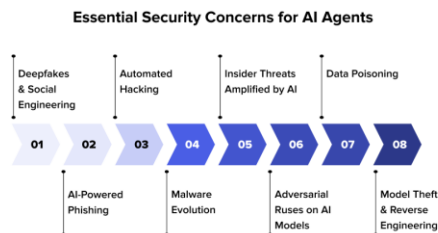


Fig. 1. AI agent Security risks

The advent of AI agents using tools has posed complex security issues since they are now capable of

independently communicating with the enterprise resources [4]. Current research has listed the presence of threats like immediate injection, overindulgence of authorizations, data leakage and unauthorized tools execution as a few of the biggest threats to autonomous systems [5]. Though scientists have worked on the approaches to identify AI-vulnerabilities, most of the approaches to the problem concentrate on the model level security, but not access control over the operations [6]. Moreover, the evaluations that are currently underway tend to focus on quantifying the performance gains without putting into account security consequences of greater autonomy [7]. For this, current literature draws attention to the necessity to establish security frameworks according to which the work of AI agents is controlled and at the same time, their decision-making possibilities are not lost [8].

Access-Control Mechanisms for Autonomous AI Systems and Enterprise Environments

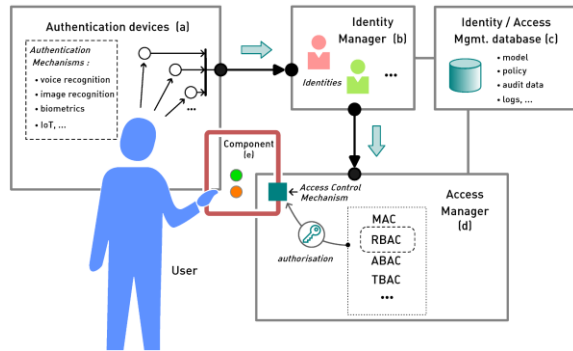


Fig. 2. Opportunities and Challenges of Artificial Intelligence Applied to Identity and Access Management in Industrial Environments

Conventional access-control techniques, such as Role-Based Access Control (RBAC), Attribute-Based Access Control (ABAC), and Policy-Based Access Control (PBAC) are structured approaches to the control of resource permissions [9]. Nevertheless, they are mostly created to be used by human operators and predictable software services, and fail to be suitable for autonomous AI agents [10]. Recent studies discuss the adaptive authorization models able to adapt to different conditions; nevertheless, a lot of the methods are based on the existing set of rules and the fixed permission systems [11]. This weakness lowers their suitability in dynamic business contexts where AI agents will have various different engagements [12]. Therefore, flexible access-control strategies to accommodate autonomous systems with task-specific and context-based permissions need to be created.

Capability-Based Least-Privilege Frameworks for Secure AI Agent Operations

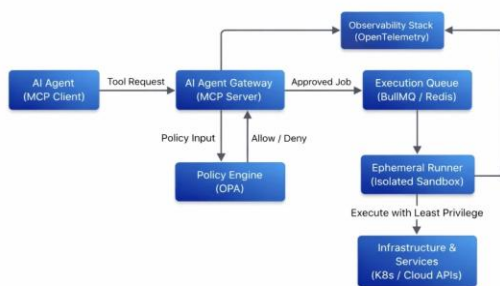


Fig.3. Least-Privilege AI Agent Gateway for Infrastructure Automation

Capability-based security presents the possibility of a new method on managing permission of AI agents by giving out the minimum of resources needed to perform distinct tasks [13]. The literature shows that the least-privilege models minimize the attack surfaces through restricting access that is not needed and barring the expansion of privileges [14]. The majority of capability-based models are however tailored towards traditional apps as opposed to intelligent agents that possess the reasoning capabilities [15]. Existing literature lacks proper instructions on how to integrate the use of this method of decision-making based on the LLM and the use of deterministic authorization mechanisms [16]. As well, there are still challenges on the generation of dynamic capability, permission validation, and continuous monitoring [17]. Thus, more research is required to institute the capability-based structures to harmonize AI autonomy, security, and the enterprise operational needs.

Policy-Based Authorization and Risk Reduction in AI Security Governance



Fig.4. Security and Policy control for AI Governance

Policy-based authorization has become an appealing tool of implementing security determination following pre-set governance principles and situational analysis [18]. Studies have shown that the use of segregation between the performance of authorization and the performance of applications enhances transparency, accountability, and compliance management [19]. Nonetheless, the current policy-based methods are typically not evaluated in the framework of autonomous AI where decisions and actions might vary dynamically [20]. Despite the focus on system governance and auditing, no studies look at the way in which policy engines can be used to govern AI agents that communicate with various enterprise tools [21].

Hence, a large research gap is in creating unifying authorization systems that involve the enforcement of regulations, the principle of least privileges, and ongoing risk evaluation of safe AI agent usage.

Research Gap

The current body of literature on AI agent security has mainly concentrated on the vulnerability-detection, the customary access-control or the governance precepts. Nonetheless, minimal efforts have been done to consider dynamic least-privilege capability control of tool-using AI agents [22]. Existing IAM strategies do not have task-focused authorization and decoupling of LLM reasoning and enforcement, necessitating dynamic enterprise AI access-control models.

III. RESEARCH METHODOLOGY

Research Philosophy

This study uses a pragmatist philosophy of applying both theoretical and practical analysis of AI agent security mechanisms. Pragmatism advocates the design of a least-privilege capability framework and verifies the effectiveness of the same by conducting experimental analysis. Its methodology facilitates incorporation of security premises, access-control schemes and machine learning-powered assessment mechanisms to gauge autonomous AI functions.

Research Approach

The study is based on a Design Science Research (DSR) methodology to create and test a framework artifact. DSR is an appropriate choice since the study supposes the development of a safe authorization scheme of tool-using AI agents and check its practical efficiency [23]. The proposed framework is tested on an enterprise-level with the help of permission analysis and machine learning models to define security patterns and forecast the levels of risk.

Research Design

A mixed evaluation plan is implemented, the development of the conceptual framework is performed and an experimental validation is conducted. The conceptual phase builds the AI agent architecture, capability permission model and policy-based authorization workflow [24]. The experimental phase is a comparison of traditional AI agents themselves, RBAC-based, and the planned dynamic least-privilege framework.

Models of machine learning, such as Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB), are used to categorize possible security risks, based on factors such as volume of permissions, sensitivity of this task, the type of resource, and frequency of unauthorized action.

The probability of risk prediction has the following form:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

Where Y represents the occurrence of security risks, X is access-control features and, β is model coefficients.

Data and Evaluation Scenario

A synthetic enterprise data is applied that includes APIs, databases, clouds, repositories, and other monitoring systems. Examples of tasks that AI agents do are code-deployment, database queries, monitoring, and incident response. Accuracy, precision, recall, F1-score and ROC-AUC are used to evaluate model performance.

Workflow Diagram

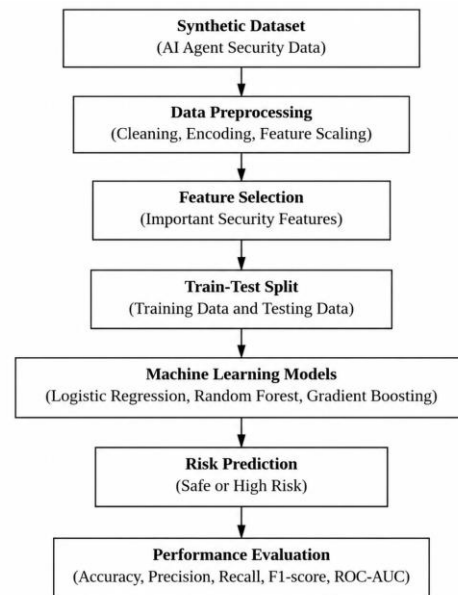


Fig. 5. Machine Learning Workflow for AI Agent Security Risk Prediction

The figure illustrates the steps in the workflow of synthetic datasets generation to preprocessing, feature selection, training of a model, risk prediction, and

performance analysis based on machine learning algorithms.

Evaluation Metrics

Permission reduction, use rate on task, preventing unauthorized action, execution time, and auditability are used as metrics of framework effectiveness. The F1-score can be computed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where accuracy is the right risk identification and accuracy the appearance capacity. These combined measures are to check security enhancement, as well as efficiency of the operational work of the proposed AI agent authorization framework.

IV. PROPOSED LEAST-PRIVILEGE CAPABILITY FRAMEWORK

Framework Overview

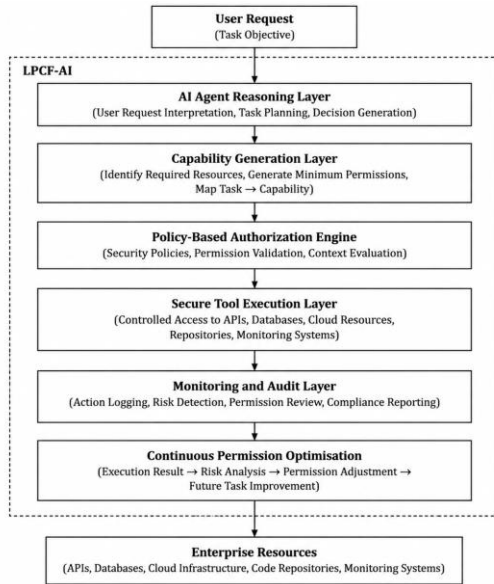


Fig. 6. Overall Architecture of Least-Privilege Capability Framework for AI Agents (LPCF-AI)

The study satisfies this need by proposing a new AI agent control model called Least- Privilege Capability Framework for AI Agents (LPCF-AI) as a means to allow both peaceful and safe coexistence between autonomous AI agents and business resources [25]. The framework isolates AI argumentation abilities and authorization choices by the introduction of autonomous ability generation, authorization implementation, safe running, and continual

observation layers. LPCF-AI in contrast to the conventional access-control methods, dynamically issues minimum permissions needed to accomplish particular tasks and minimize the risks of over-access.

AI Agent Reasoning Layer

The AI Agent Reasoning Layer deals with interpreting the user requests, planning and producing operational decisions. Interpretation of objectives and appropriate actions is done with Large Language Models (LLM); nevertheless, permissions and access to resources are not directly controlled by the LLM [26]. This isolation makes unauthorized actions due to autonomous reasoning errors unachievable, and security decisions that are under deterministic authorization mechanisms.

Capability Generation Layer

The Capability Generation Layer transforms planned AI tasks into particular access requirements. It determines needed enterprise resources, creates least necessary permissions and maps tasks to suitable capabilities. As an example, a repository read, CI/CD pipeline read, and deployment execute permission can be needed by an application deployment task [27]. However, unnecessary privileges like modifications on user accounts or accessing financial databases remain restricted.

$$\begin{aligned} \text{Capability} = & \text{Task Requirement} \\ & + \text{Resource Sensitivity} \\ & + \text{Minimum Permission} \end{aligned}$$

The strategy compares to make sure that things are created based on the need to operate instead of pre-defined general operations.

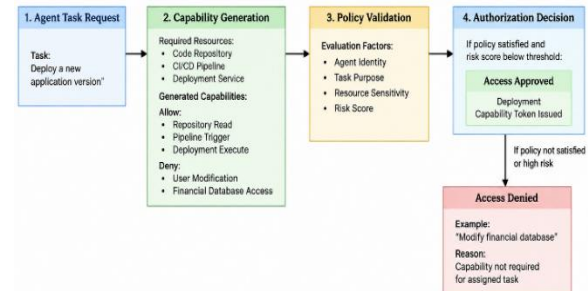


Fig.7. End-to-End Capability Generation and Authorization Workflow

This figure shows an end-to-end authorization process that reflects how AWS Storage Gateway works with

LPCF-AI to authorize end to end access to a storage gateway. AI-agent task is analyzed and minimum capabilities are generated from this analysis, regarding required enterprise resource. These capabilities are vetted against security policies and within the context of risk factors. Capability tokens are issued to approved requests and denied to unnecessary or high-risk requests, showing responsibility for request privilege by least privilege. Capability tokens are created for approved requests, and denied for unnecessary or high-risk requests, with responsibility for request privilege by least privilege.

Permission Generation Pseudocode

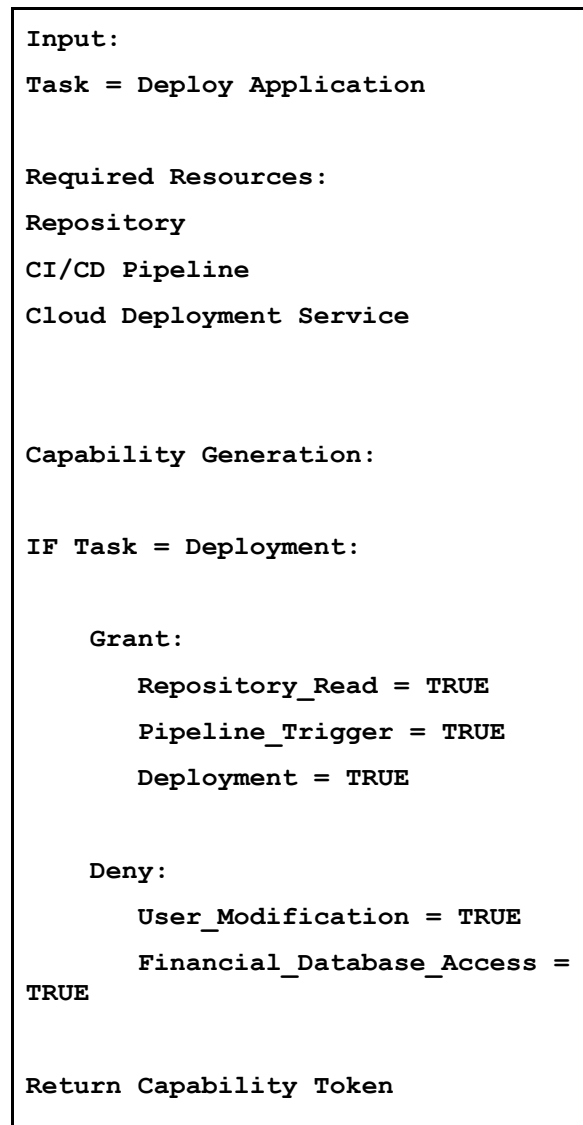


Fig. 8. Permission Generation Pseudocode

LPCF-AI takes operational tasks and transforms them into minimum required capabilities, which is shown by the permission generation pseudocode [28]. It grants minimal permissions on executing its tasks, but explicitly refuses non-essential permissions, which provides secure task-specific investigations of AI agents.

Policy-Based Authorization Engine

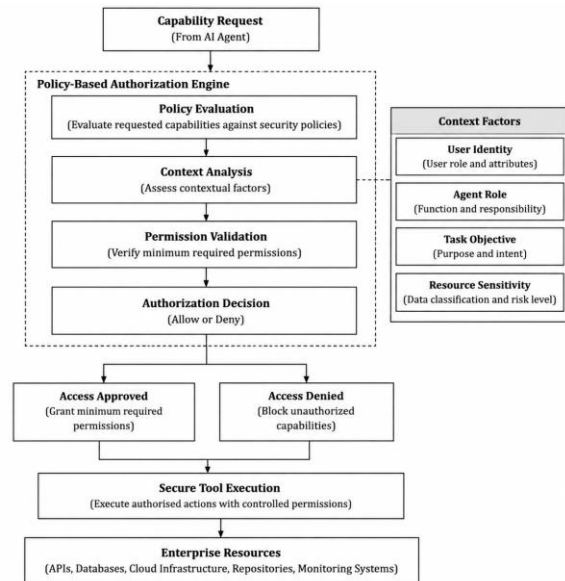
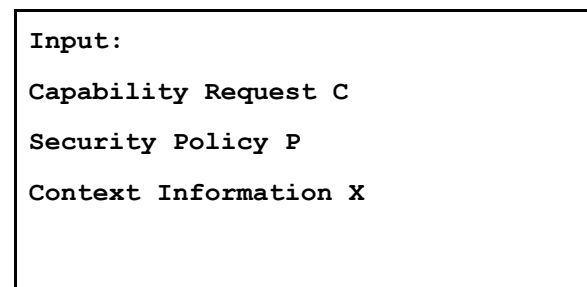


Fig. 9. Context-Aware Policy Authorization Workflow

The Policy-Based Authorization Engine is a key security enforcement engine. It checks the capabilities generated with stipulated security policies and permits AI agents to reach into enterprise assets [29]. The engine considers contextual information such as the user is, the agent position, the task purpose, and sensitivity of resources. This also makes sure that any AI action is approved in terms of organizational security needs and compliance.

Authorization Decision Algorithm



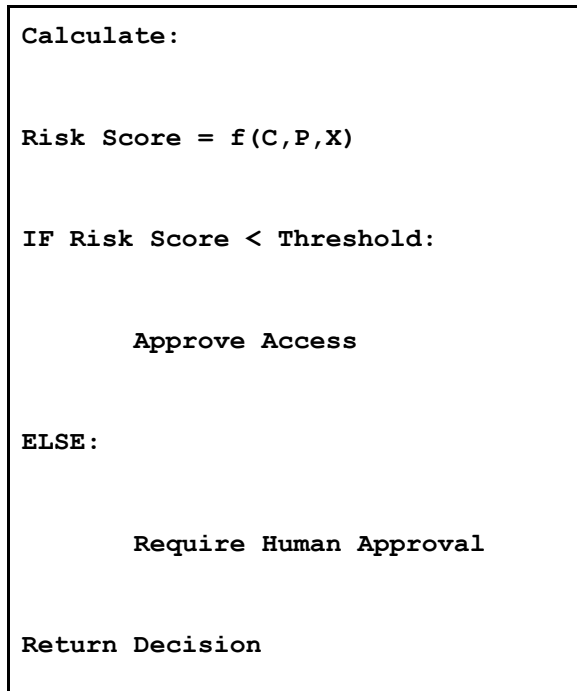


Fig. 10. Authorization Decision Algorithm

This pseudocode illustrates the process of authorization decision could be conducted as it calculates the contextual risk scores and authentication of the capability requests according to security policies in order to allow controlled access, reject requests, or request human intervention.

Secure Tool Execution Layer

The Secure Tool Execution Layer is a controlled engagement between AI agents and business tools [30]. It enforces technical controls both within APIs, databases, and code repositories and in the cloud. Only valid policy results in accessions, thwarting unauthorized modification and over access of data, and possible privilege elevation during a self-governing operation.

Monitoring and Audit Layer

The Monitoring and Audit Layer offers real-time insight into the actions of AI agents by logging

actions, identifying risk, review permissions, and compliance reporting. All acts being performed are logged to assist with accountability, forensic investigation and governance. This layer enhances transparency as organizations are able to gain insight into the uses of assigned capabilities by AI agents.

Continuous Permission Optimization

Continuous Permission Optimization layer allows adaptive security management based on feedback. The outcomes of execution of AI agents are analyzed to determine possible risks, unwarranted privileges, and authorization improvements. Future tasks are dynamically provided on the basis of the results of risk assessment. The process enables LPCF-AI to sustain secure operations and take into consideration the autonomy of agents, their efficiency, and corporate governance demands.

V. FRAMEWORK EVALUATION AND RESULTS

Experimental Scenario

An AI DevOps agent by way of an artificial enterprise is used in the proposed Least-Privilege Capability Framework of AI Agents (LPCF-AI) to evaluate how effectively the agents operated and could securely operate in the proposed framework. The experimental scenario consisted of three big activity types such as code review, application deployment and infrastructure monitoring. All these activities demanded censored interactions with enterprise resources such as code repositories, CI/CD pipelines, cloud infrastructure, database and systems, and monitoring.

Security Scenario-Based Evaluation

LPCF-AI is tested with typical AI-agent security scenarios to assess its practical security effectiveness.

TABLE 2. SECURITY SCENARIO-BASED EVALUATION

Security Scenario	Threat Description	Traditional AI Agent Behaviour	LPCF-AI Response
Excessive Permission Request	Agent requests unnecessary administrative or database privileges	May allow excessive access due to broad permissions	Generates minimum required capabilities and rejects unnecessary privileges

Unauthorized Resource Access	Agent attempts access to restricted enterprise resources	Access may occur if excessive permissions exist	Policy engine denies requests that violate capability policies
Prompt Injection Attempt	Malicious instruction attempts to influence agent actions	Agent may execute unintended operations	Authorization layer validates actions independently from LLM reasoning
Privilege Escalation Attempt	Agent requests higher-level permissions than required	Increased risk of unauthorized modification	Capability generation blocks non-task-related permissions

As illustrated in the scenario evaluation, LPCF-AI minimizes security risks with AI agents by decoupling, which is reasoning from authorization enforcement. All actions are handled through capability generation and validation of policies, potentially eliminating the chance of giving too much privileges, unauthorized access and privilege escalation attempts directly to enterprise resources by autonomous agents.

The assessment considered the effects of various authorization strategies on prediction of security risks, permission and performance of the operation. The logistic regression, the random forest, and the gradient boosting machine learning model were used to classify the AI-agent security risks according to the volume of permission, sensitivity of the resources, historical risk behavior, failed to gain access, frequency of access, and authorship. The comparative ML analysis indicated that security-risk prediction is still a difficult task as there are many relationships between the autonomous agent behavior and enterprise resource access patterns that are hard to figure out.

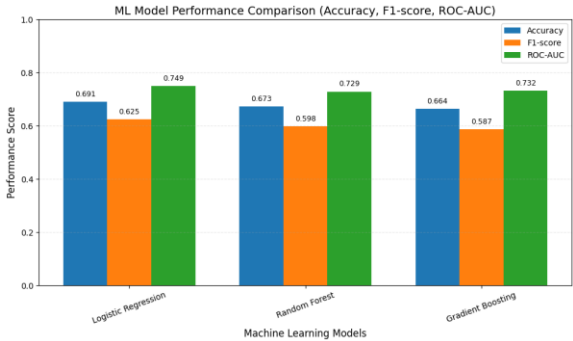


Fig. 11. ML Model Performance Comparison (Accuracy, F1-score, ROC-AUC)

This figure shows the comparison of the performance of the devised models. The best accuracy (0.691), F1-score (0.625) and ROC-AUC (0.749) are achieved through logistic regression, which has been shown to have greater classification strength than Random Forest and Gradient boosting.

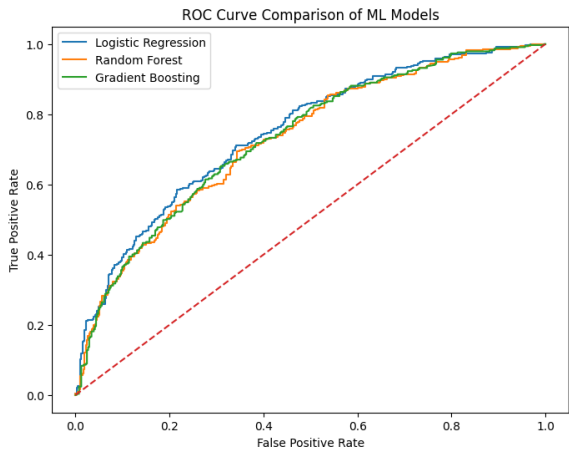


Fig. 12. ROC Curve Comparison of ML Models

This figure shows that this or that model is capable of discrimination. Logistic Regression though had the highest ROC-AUC of 0.749 when compared to random Forest and Gradient Boosting which had 0.729 and 0.732 respectively. This implies that the models have realistic predictive performance in the detection of security risks in autonomous environments of AI-agents.

Comparative Evaluation

For comparing the offered LPCF-AI framework with the traditional full-permission agents and the RBAC based agents, the frameworks were tested in terms of the improvement in security.

TABLE 3. COMPARATIVE EVALUATION

Model	Access Level	Security Risk	Task Performance
<i>Full Permission Agent</i>	<i>High</i>	<i>High</i>	<i>High</i>
<i>RBAC Agent</i>	<i>Medium</i>	<i>Medium</i>	<i>Medium</i>
<i>LPCF-AI Agent</i>	<i>Minimum</i>	<i>Low</i>	<i>High</i>

The traditional full permission AI agents can offer the highest flexibility in their activities, however become a major source of security vulnerability since too much power leads to the increased possibility of unauthorized access and misuse of resources. The RBAC-based agents offer a better control; nevertheless, predefined roles might still surpass the tasks requirements. LPCF-AI intelligently allocates only necessary capabilities, which allow safe autonomous work without decreasing the performance of tasks.

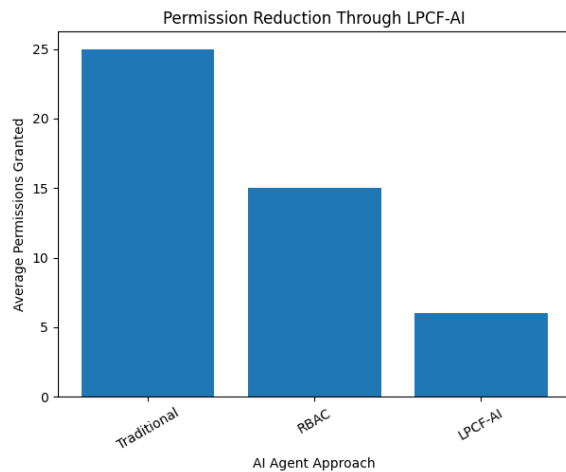


Fig. 13. Permission Reduction Through LPCF-AI

LPCF-AI has permission optimization power, as can be seen in this upper figure. It is analyzed that the traditional AI agents will need about 25 permissions, RBAC agents need 15 permissions, and LPCF-AI agents need only 6 permissions. This is a 76% decrease

over the conventional methods of authorization and shows enhanced least privilege implementation.

Security Improvement Analysis

The analysis confirms that LPCF-AI offers enterprise AI-agent security by reducing the needless permissions and preventing privilege escalation as well as improving auditability. The architecture eliminates overreacting by prohibiting the space between LLDM reasoning and deterministic authorization measures, which implies that AI-generated decisions cannot change enterprise resources directly, without policy validation.

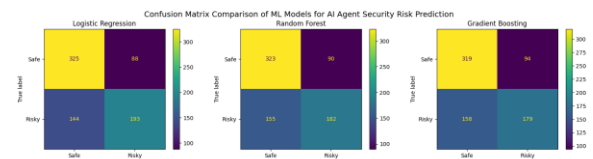


Fig. 14. Confusion Matrix Comparison of ML Models for AI Agent Security Risk Prediction

This figure is also analyzed to determine the reliability of security-risk classification. Eighty-eight incorrectly classified safe and 144 incorrectly classified risky cases Logistic Regression correctly classified 325 safe and 193 risky cases, and falsely classified 88 and 144 false negative cases, respectively. Random Forest and Gradient Boosting showed somewhat higher classification errors, which have relatively less prediction balance.

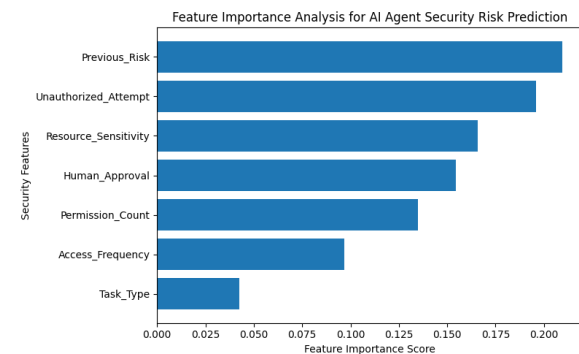


Fig.15. Feature Importance Analysis for AI Agent Security Risk Prediction

This figure defines the significant contributors to security risks by AI-agents. The scores of Importance were highest in Previous Risk (0.209) then Unauthorized Attempt (0.196) then Resource Sensitivity (0.166). These results suggest that

historical behavior, unauthorized activities and resource criticality are noteworthy factors in adaptive authorization decisions.

Performance Evaluation

Value Completion accuracy, response efficiency and tool performance were the measures used to determine the performance of LPCF-AI. The ML analysis revealed that the security-risk is well-classified, and the predictive performance of Logistic Regression is the best. Though LPCF-AI is also operationally efficient due to the validation of only task-specific capabilities prior to execution, the introduction of extra authorization steps creates some processing overhead, although that is minor.

The framework has managed to create a balance between security and autonomy by enabling AI agents to perform authorized tasks and preventing unjustified access. Repetitive observation, authorization, and audit logging will also enhance governance to ensure there is visibility of the activities of the AI-agents. Altogether, the analysis shows that LPCF-AI offers a safe and scalable model to manage AI-agents in an enterprise efficiently mitigating security risk without decreasing operational performances.

VI. DISCUSSION

The analysis supports the idea that LPCF-AI is an effective balance of AI autonomy and enterprise

security as it allows agents to execute tasks with the least required permissions. The framework minimized the number of requirements of 25 permissions to 6 with a 76% drop with high task performance. Logistic Regression had an accuracy of 0.691 and a ROC-AUC of 0.749, which is practical to classify risks. But the complexity of permission management, random agent behavior and governance issues are still a major concern. Consequently, to have secure, transparent, and reliable enterprise adoption of autonomous AI agents, adaptive authorization policies, constant monitoring, and human control are vital.

Practical Enterprise Integration

LPCF-AI is not to be aimed at replacing existing Identity and Access Management (IAM) systems, however instead it can be used as an additional authorization layer within enterprise security infrastructures. LPCF-AI can dynamically create capabilities for AI agents and enforce authorizations based on security policies, resource sensitivity, and operational risk. Existing IAM solutions can handle the authentication of AI agents; LPCF-AI can dynamically create capabilities, and enforce authorizations based on security policies, resource sensitivity and operational risks. With this integration, autonomous AI agents can be used securely across enterprise applications.

TABLE 4. PRACTICAL ENTERPRISE INTEGRATION

Enterprise Component	Integration Approach
Identity and Access Management (IAM)	LPCF-AI integrates with existing identity providers to authenticate AI agents before capability generation and authorization decisions.
Cloud Security Platforms	Capability-based policies can regulate AI-agent access to cloud APIs, containers, deployment services, and infrastructure resources.
Policy Management Systems	Existing enterprise policy engines can provide governance rules and security constraints for capability validation and approval decisions.
Security Monitoring Platforms	LPCF-AI audit records can integrate with SIEM platforms to support continuous monitoring, threat detection, and compliance reporting.

This integration model enables LPCF-AI to establish a secure demarcation zone between an AI agent and enterprise resources to ensure a controlled environment. The framework can allow for the different functions of autonomy; however, it also

contain the validation of each action that would be requested based on the nature of the task, organizational procedure and risk assessment early in the process, and prior to execution.

VII. ETHICAL AND GOVERNANCE CONSIDERATIONS

LPCF-AI is a beneficial approach to the appropriate use of AI as there is a clear authorization, human influence, and accountability system. The framework is in line with the principles of NIST AI Risk Management Framework, ISO/IEC 27001, and Zero Trust Architecture as it ensures that risks are managed, secure access control, continuous verification, and governance of autonomous AI processes.

VIII. CONCLUSION AND FUTURE RESEARCH

Conclusion

This study illustrates the need to implement powerful governance systems to tool-using AI agents within businesses. The LPCF-AI framework presents capability-based least-privilege authorization where the basis of the to-be-motivated security controls is separated using the reasoning of the LLM. This solution minimizes unwarranted permissions, enhances auditability, and facilitates sound AI adoption without impacting performance.

Future Direction

The validation of LPCF-AI should be conducted in future research by real-life enterprise deployments and in different operating conditions. Future research can investigate autonomous policy learning, generation of adaptive capabilities and compatibility with the existing IAM systems. Moreover, explainable authorization mechanisms and ongoing risk evaluation model investigations have the potential to enhance scalability, transparency, and safe integration of AI agents.

IX. REFERENCES

- [1]Guan, S., Lin, F., Li, J., Wang, J. and Wang, F.Y., 2024. Parallel Financial Systems: Towards Governable and Sustainable Intelligent Financial Services. *Journal of Cyber-Physical-Social Intelligence*, 3(1), pp.9-9.
- [2]Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., Qin, Z. and Ren, K., 2024. Privacyasst: Safeguarding user privacy in tool-using large language model agents. *IEEE Transactions on Dependable and Secure Computing*, 21(6), pp.5242-5258.
- [3]Tang, X., Jin, Q., Zhu, K., Yuan, T., Zhang, Y., Zhou, W., Qu, M., Zhao, Y., Tang, J., Zhang, Z. and Cohan, A., 2024. Prioritizing safeguarding over autonomy: Risks of llm agents for science. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- [4]Anis, F. and Hammoudeh, M., 2024, December. Weaponizing AI in Cyberattacks A Comparative Study of AI powered Tools for Offensive Security. In *Proceedings of the 8th International Conference on Future Networks & Distributed Systems* (pp. 283-290).
- [5]Jaber, A. and Fritsch, L., 2022, October. Towards ai-powered cybersecurity attack modeling with simulation tools: Review of attack simulators. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 249-257). Cham: Springer International Publishing.
- [6]Shahid, J., Muhammad, Z., Iqbal, Z., Khan, M.S., Amer, Y. and Si, W., 2022, March. Sat: Integrated multi-agent blackbox security assessment tool using machine learning. In *2022 2nd International Conference on Artificial Intelligence (ICAI)* (pp. 105-111). IEEE.
- [7]Santoso, F. and Finn, A., 2023. An in-depth examination of artificial intelligence-enhanced cybersecurity in robotics, autonomous systems, and critical infrastructures. *IEEE Transactions on Services Computing*, 17(3), pp.1293-1310.
- [8]Hamad, M. and Steinhurst, S., 2023, November. Security challenges in autonomous systems design. In *International Conference on Computational Technologies and Electronics* (pp. 142-154). Cham: Springer Nature Switzerland.
- [9]Wang, J., Jiao, Z., Chen, J., Hou, X., Yang, T. and Lan, D., 2023. Blockchain-aided secure access control for UAV computing networks. *IEEE Transactions on Network Science and Engineering*, 11(6), pp.5267-5279.
- [10]Sadeghi, M., Sartor, L. and Rossi, M., 2022. A semantic-based access control approach for systems of systems. *ACM SIGAPP Applied Computing Review*, 21(4), pp.5-19.
- [11]CHERIF, A.N., YOUSSEFI, M., EN-NAIMANI, Z., TADLAOUI, A., SOULAMI, M. and

- BOUATTANE, O., 2024. CQRS and Blockchain with Zero-Knowledge Proofs for Secure Multi-Agent Decision-Making. *International Journal of Advanced Computer Science & Applications*, 15(11), p.892.
- [12]Xiao, S., Ye, Y., Kanwal, N., Newe, T. and Lee, B., 2022. SoK: Context and risk aware access control for zero trust systems. *Security and Communication Networks*, 2022(1), p.7026779.
- [13]James, M., Newe, T., O'Shea, D. and O'Mahony, G.D., 2024, June. Authentication and authorization in zero trust IoT: A survey. In *2024 35th Irish Signals and Systems Conference (ISSC)* (pp. 1-7). IEEE.
- [14]Burhan, M., Alam, H., Arsalan, A., Rehman, R.A., Anwar, M., Faheem, M. and Ashraf, M.W., 2023. A comprehensive survey on the cooperation of fog computing paradigm-based IoT applications: layered architecture, real-time security issues, and solutions. *IEEE Access*, 11, pp.73303-73329.
- [15]Zambare, P. and Liu, Y., 2023, October. Understanding security challenges and defending access control models for Cloud-based Internet of Things network. In *IFIP International Internet of Things Conference* (pp. 179-197). Cham: Springer Nature Switzerland.
- [16]Islam, M.Z. and Dhanekula, A., 2023. Measuring the Security Impact of Zero Trust Access Controls: A Mixed-Methods Study of Identity-Based Policies (Cisco ISE+ AD) and Incident Reduction. *American Journal of Data Science and Analytics*, 4(06), pp.01-42.
- [17]Chaisiri, A. and Boonmee, K., 2024. Authentication and Authorization Mechanisms in Secure Systems: Their Impact on Information Assurance and Access Control. *Transactions on Embedded Systems, Real-Time Computing, and Applications*, 14(6), pp.1-14.
- [18]Jin, Z., Xing, L., Fang, Y., Jia, Y., Yuan, B. and Liu, Q., 2022, November. P-verifier: Understanding and mitigating security risks in cloud-based IoT access policies. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1647-1661).
- [19]Choudhary, A.R., 2023. Enhancing cybersecurity using a new dynamic approach to authentication and authorization. *Issues in Information Systems*, 24(2), p.22.
- [20]Dabra, M., Sharma, S., Kumar, S. and Min, H., 2024. An improved finegrained ciphertext policy based temporary keyword search on encrypted data for secure cloud storage. *Scientific Reports*, 14(1), p.5264.
- [21]Jangala, V.K., 2024. Authentication and authorization mechanisms in Java-based systems. *International Journal of Contemporary Research in Multidisciplinary*, 3(1), pp.277-284.
- [22]Vadlapatla, S., 2024. Schema-Grounded Agentic AI for Regulatory-Compliant Data Access: Bridging Natural Language and Enterprise Data Governance in Financial Services. *International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)*, 7(4), pp.10861-10868.
- [23]Meesala, L.K., 2023. A layered security framework for enterprise operations in the Generative AI and Agentic AI era in regulated cloud environments. *International Journal of Future Innovative Science and Technology (IJFIST)*, 6(6), p.11760.
- [24]Feretzakis, G. and Verykios, V.S., 2024. Trustworthy AI: Securing sensitive data in large language models. *Ai*, 5(4), pp.2773-2800.
- [25]Kang, H., Liu, G., Wang, Q., Meng, L. and Liu, J., 2023. Theory and application of zero trust security: A brief survey. *Entropy*, 25(12), p.1595.
- [26]Jeffrey, N., Tan, Q. and Villar, J.R., 2023. A review of anomaly detection strategies to detect threats to cyber-physical systems. *Electronics*, 12(15), p.3283.
- [27]Sai, S., Yashvardhan, U., Chamola, V. and Sikdar, B., 2024. Generative AI for cyber security: Analyzing the potential of ChatGPT, DALL-E, and other models for enhancing the security space. *IEEE access*, 12, pp.53497-53516.
- [28]Ragothaman, K., Wang, Y., Rimal, B. and Lawrence, M., 2023. Access control for IoT: A survey of existing research, dynamic policies and future directions. *Sensors*, 23(4), p.1805.
- [29]Tariq, U., Ahmed, I., Bashir, A.K. and Shaukat, K., 2023. A critical cybersecurity analysis and future

research directions for the internet of things: a comprehensive review. *sensors*, 23(8), p.4117.

[30]Awan, S.M., Azad, M.A., Arshad, J., Waheed, U. and Sharif, T., 2023. A blockchain-inspired attribute-based zero-trust access control model for IoT. *Information*, 14(2), p.129.