

Policy-Driven Access Control for Secure Deployment of Autonomous AI Agents in Enterprise Environments

Harish Gaggar*

Staff Software Site Reliability Engineer

* Corresponding Author Email: mrgaggarg@gmail.com ORCID: 0000-0002-5247-782X

Article Info:

DOI: 10.22399/ijcesn.5497
Received : 01 November 2025
Revised : 24 December 2025
Accepted : 26 December 2025

Keywords

autonomous AI agents,
policy-driven access control,
enterprise security,
least privilege,
role-based access control,
human-in-the-loop

Abstract:

The growing adoption of autonomous artificial intelligence agents in enterprise environments introduces significant security challenges because such agents can independently access data, invoke tools, execute workflows, and interact with multiple organizational systems. This study evaluates a policy-driven access-control framework designed to regulate autonomous agent behavior through least-privilege authorization, contextual policy evaluation, risk-adaptive enforcement, human approval, and comprehensive audit logging. A controlled enterprise simulation was developed to compare minimal control, conventional role-based access control, and policy-driven access control across security and operational performance indicators. The policy-driven framework reduced privilege exposure ratio to 0.17 ± 0.04 , unauthorized action rate to $2.65 \pm 2.25\%$, and privilege-escalation success to $1.46 \pm 1.37\%$, while increasing malicious-action blocking to $97.61 \pm 1.49\%$, policy-enforcement accuracy to $96.05 \pm 1.67\%$, and audit completeness to $98.92 \pm 0.86\%$. Privilege exposure was strongly positively correlated with unauthorized action rate ($r = 0.91$), and regression analysis identified privilege exposure as the strongest positive predictor of unauthorized behavior ($\beta = 0.507, p < 0.001$), whereas policy strictness exerted a significant negative effect ($\beta = -0.429, p < 0.001$). Although policy-driven enforcement increased authorization latency to 68.86 ± 9.41 ms and slightly reduced legitimate task completion to $90.92 \pm 2.82\%$, the overall security gains substantially outweighed the operational overhead. The findings demonstrate that policy-driven access control can provide an effective and scalable security foundation for autonomous AI agents by balancing operational autonomy with enterprise governance, accountability, and risk containment.

1. Introduction

Emergence of autonomous AI agents in enterprise systems

Artificial intelligence has evolved from conventional rule-based decision support and predictive analytics toward increasingly autonomous systems capable of planning, reasoning, communicating, and executing tasks with limited human intervention (Acharya et al., 2025). Autonomous AI agents represent an important stage in this evolution because they can interact with enterprise applications, retrieve and process organizational data, invoke external tools, execute workflows, and dynamically adapt their actions according to changing operational conditions (Koubaa, 2025). Enterprises are consequently exploring agent-based architectures

for customer service, software engineering, business intelligence, cybersecurity, financial operations, supply-chain management, and administrative automation. Unlike conventional applications, however, autonomous agents may operate across multiple interconnected systems and make sequential decisions that were not individually specified by developers (Hughes et al., 2025). This operational autonomy creates substantial opportunities for efficiency and scalability while simultaneously introducing new security concerns regarding how agents obtain, retain, and exercise access privileges within enterprise environments.

Security challenges of agent autonomy

The security requirements of autonomous AI agents differ considerably from those of conventional enterprise software. Traditional applications

generally operate according to predefined workflows and relatively stable authorization structures, whereas AI agents can dynamically select tools, formulate intermediate objectives, access heterogeneous resources, and modify their execution paths in response to contextual information (Bandi et al., 2025). An agent authorized to query a database, send messages, modify cloud resources, or trigger business processes may therefore possess the technical capacity to perform actions extending beyond the original intention of a user request (Igwe-Nmaju, 2024). Prompt injection, compromised contextual information, incorrect reasoning, excessive permissions, malicious tool responses, and agent-to-agent instruction propagation can further influence autonomous behavior (Adabara et al., 2025). Consequently, authentication alone cannot guarantee secure agent deployment. Enterprises require mechanisms capable of continuously determining not merely whether an agent is authenticated, but whether a particular action is permissible within its current identity, purpose, resource scope, operational context, and organizational policy (South et al., 2025).

Limitations of conventional access-control models

Enterprise security has historically relied on mechanisms such as role-based access control, attribute-based access control, mandatory access control, and discretionary authorization. These approaches remain fundamental, yet autonomous AI introduces operational conditions for which static authorization structures may be insufficient (Acharya et al., 2025). Assigning a broad organizational role to an AI agent can generate excessive privileges because the same agent may perform multiple tasks requiring different levels of authorization. Similarly, granting persistent credentials for convenience can violate the principle of least privilege and enlarge the potential attack surface (Kalejaiye & Shonubi, 2025). Autonomous workflows may also cross application boundaries, creating situations in which permissions appropriate in one system become inappropriate when combined with capabilities available elsewhere. Access decisions must therefore account for the requested operation, initiating user, agent identity, data sensitivity, environmental context, workflow purpose, execution history, and potential consequences of the proposed action (Kalejaiye & Shonubi, 2025). These requirements encourage a transition from predominantly static permission assignment toward policy-driven and context-sensitive authorization.

Policy-driven access control for AI agents

Policy-driven access control provides a systematic mechanism for translating organizational security

requirements into machine-enforceable rules governing agent behavior (Saidur & Khan, 2023). Rather than granting unrestricted tool or resource access after authentication, policies can evaluate individual requests according to predefined conditions before execution. Such policies may specify which resources an agent can access, which operations it can perform, under whose authority it operates, how long permissions remain valid, and under what circumstances additional approval is required (Qiu et al., 2020). They can also enforce least-privilege access, separation of duties, temporary credentials, purpose limitation, data classification requirements, and restrictions on high-impact actions. When policies are evaluated at runtime, authorization can become adaptive to the context surrounding each action. This capability is particularly important for autonomous agents because their execution trajectories may evolve dynamically and cannot always be completely predicted during system design.

Governance, accountability and enterprise compliance

Secure autonomous AI deployment also requires strong governance and accountability mechanisms. Enterprises operate under internal policies, contractual obligations, industry standards, and regulatory requirements concerning privacy, cybersecurity, data management, and operational risk (Wylde et al., 2022). Autonomous agents must therefore generate sufficient evidence to establish who initiated an operation, which agent performed it, what resources were accessed, which policy authorized the action, and what outcome resulted (Dennis et al., 2016). Policy enforcement combined with immutable audit trails, identity management, authorization logs, human approval checkpoints, and continuous monitoring can improve traceability throughout the agent lifecycle (Huang & Hughes, 2025). Such controls are especially important when agents perform financially significant, privacy-sensitive, security-critical, or irreversible operations. Policy-driven architectures can therefore connect technical authorization mechanisms with broader organizational governance by making security requirements explicit, enforceable, observable, and auditable.

Need for a secure deployment framework

Despite rapid adoption of generative and agentic AI technologies, a significant gap remains between demonstrating autonomous capabilities and deploying them securely within production enterprise environments. Many experimental agent systems emphasize task completion, reasoning performance, and tool-use efficiency while treating authorization as a secondary infrastructure concern (Krishnan, 2025). Enterprise deployment requires a

different perspective in which security boundaries are integrated directly into agent execution (Huang & Hughes, 2025). A secure framework should distinguish user authority from agent capability, provide task-specific permissions, evaluate access continuously, restrict lateral movement across resources, protect sensitive information, and provide escalation mechanisms when policy conditions are uncertain or violated. Such an architecture should preserve useful agent autonomy while preventing autonomy from becoming unrestricted authority.

Research focus and contribution

Against this background, the present study investigates policy-driven access control as a security foundation for deploying autonomous AI agents in enterprise environments. The research focuses on designing an authorization architecture in which agent actions are evaluated against explicit organizational policies before access to protected resources or execution of privileged operations is permitted. Particular attention is given to agent identity, least-privilege authorization, contextual policy evaluation, dynamic permission boundaries, human oversight, auditability, and enforcement across heterogeneous enterprise resources. The study further examines how policy-driven controls can reduce unauthorized or excessive agent actions while maintaining the operational flexibility required for autonomous workflows. By integrating enterprise access-control principles with the distinctive execution characteristics of agentic AI, the research seeks to establish a practical framework for balancing autonomy, security, compliance, and operational efficiency. The proposed perspective contributes toward the development of trustworthy enterprise AI systems in which autonomous capabilities can be scaled without weakening institutional control over sensitive data, critical applications, and consequential digital actions.

2. Methodology

Research design

The study adopted an experimental systems-evaluation design to assess the effectiveness of policy-driven access control for secure deployment of autonomous AI agents in enterprise environments. A controlled enterprise simulation environment was developed in which autonomous agents were permitted to interact with multiple organizational resources, including structured databases, document repositories, application programming interfaces, messaging systems, cloud services, and administrative tools. The experimental framework was designed to compare autonomous-

agent behavior under different access-control configurations and determine how effectively policy enforcement could reduce unauthorized actions without excessively restricting legitimate task execution. The analysis considered both security effectiveness and operational performance because an enterprise access-control mechanism must simultaneously minimize policy violations, limit privilege exposure, maintain task completion capability, and control authorization overhead.

Enterprise agent architecture

The experimental architecture consisted of five major components: the user layer, autonomous AI agent layer, policy decision point, policy enforcement point, and enterprise resource layer. The user layer generated task requests and supplied user identity, organizational role, and task context. Autonomous agents interpreted requests, developed execution plans, selected tools, and generated resource-access requests. Every privileged operation was intercepted by the policy enforcement point before execution. The policy decision point evaluated each request using agent identity, user authority, requested resource, intended operation, contextual attributes, sensitivity level, risk score, and current workflow state. Access was granted, denied, conditionally permitted, or escalated for human approval according to predefined security policies. All decisions and subsequent agent actions were recorded within an audit-log repository to support traceability and post-execution analysis.

Agent identity and authorization parameters

Each autonomous agent was assigned a unique agent identifier and a functional role defining its permitted operational domain. Agent-related variables included agent identity, assigned role, capability profile, authentication status, trust score, session duration, delegated authority, active permissions, and previous policy-violation history. User-related authorization parameters included user identity, organizational role, security clearance level, department, resource ownership, and authorization scope. Delegated authority was represented as the subset of user permissions explicitly transferred to the agent for completing a specific task. This distinction was used to prevent an agent from automatically inheriting the entire privilege set of the initiating user.

Authorization decisions were therefore represented as a function of multiple attributes:

$$A = f(U, G, R, O, C, S, P, T)$$

where A represents the authorization decision, U represents user attributes, G represents agent attributes, R represents the requested resource, O represents the requested operation, C represents

contextual conditions, S represents resource sensitivity, P represents applicable policy rules, and T represents temporal constraints.

Policy variables and enforcement rules

The policy framework incorporated role-based, attribute-based, context-aware, and risk-adaptive access-control principles. Major policy variables included authorized resource category, permissible action type, data-classification level, permission duration, geographic or network context, time-of-access restriction, workflow purpose, approval requirement, action reversibility, transaction value, and estimated action risk. Policies were created for read, write, modify, delete, execute, transfer, and administrative operations.

Resource sensitivity was classified into four categories: public, internal, confidential, and highly restricted. Action-risk levels were classified as low, moderate, high, and critical. Low-risk operations, such as reading non-sensitive internal documentation, could be automatically permitted if identity and role requirements were satisfied. Moderate-risk operations required additional contextual validation. High-risk operations required restricted temporary privileges or explicit approval, whereas critical operations, such as deleting production resources or transferring highly sensitive information, required human authorization. Policy enforcement followed the principle of default denial, whereby requests were rejected unless an applicable policy explicitly permitted the action.

Least-privilege and temporary permission control

Least-privilege effectiveness was evaluated by dynamically limiting the permissions available to each agent according to the current task. Instead of assigning permanent broad permissions, task-specific access tokens were generated with restricted resource scope, permitted actions, and validity duration. Three permission configurations were evaluated: broad persistent privilege, static role-based privilege, and dynamic policy-driven least privilege. Permission scope was measured as the proportion of accessible operations relative to the total operations technically available within the environment.

The privilege exposure ratio was calculated as:

$$PER = P_a / P_t$$

where PER is the privilege exposure ratio, P_a represents the number of active permissions assigned to the agent, and P_t represents the total number of permissions available within the enterprise environment. Lower PER values indicated stronger adherence to least-privilege principles.

Autonomous task scenarios

A series of enterprise task scenarios was developed to evaluate normal and adversarial agent behavior. Legitimate scenarios included retrieving business records, updating customer information, generating analytical reports, sending approved messages, invoking authorized APIs, and modifying permitted workflow parameters. Security-sensitive scenarios included requests to access unauthorized datasets, retrieve information outside the assigned department, modify protected configurations, export confidential information, execute high-risk administrative commands, and reuse credentials beyond their authorized lifetime.

Adversarial scenarios were also introduced to test policy robustness. These included prompt-injection instructions embedded within retrieved documents, malicious tool responses requesting additional privileges, cross-resource privilege escalation, indirect requests to access confidential information, conflicting user and system instructions, and attempts to execute unauthorized actions through chained tool calls. Each scenario was executed repeatedly under comparable conditions across the evaluated access-control configurations.

Experimental variables

The primary independent variable was access-control configuration, represented by unrestricted or minimally restricted access, conventional role-based access control, and the proposed policy-driven access-control framework. Additional independent variables included resource sensitivity, action-risk category, workflow complexity, number of tool calls, agent autonomy level, and presence or absence of adversarial input.

Dependent security variables included unauthorized action rate, policy-violation rate, privilege-escalation success rate, sensitive-data exposure rate, blocked malicious action rate, approval-escalation rate, and policy-enforcement accuracy. Operational variables included legitimate task completion rate, average authorization latency, total task-execution time, number of access-control interruptions, agent retry frequency, and computational overhead.

Security evaluation metrics

Unauthorized action rate was calculated as:

$$UAR = N_u / N_a \times 100$$

where N_u represents unauthorized actions successfully executed and N_a represents the total number of attempted actions.

Policy enforcement accuracy was calculated as:

$$PEA = (TP + TN) / (TP + TN + FP + FN) \times 100$$

where TP represents malicious or unauthorized requests correctly blocked, TN represents legitimate requests correctly permitted, FP represents legitimate requests incorrectly blocked,

and FN represents unauthorized requests incorrectly permitted.

The malicious action blocking rate was calculated as:

$$\text{MABR} = N_b / N_m \times 100$$

where N_b represents malicious actions successfully blocked and N_m represents the total malicious actions attempted.

The false denial rate was determined as:

$$\text{FDR} = \text{FP} / (\text{FP} + \text{TN}) \times 100$$

to measure how frequently legitimate requests were unnecessarily rejected.

The privilege-escalation success rate was calculated as:

$$\text{PESR} = N_e / N_p \times 100$$

where N_e represents successful privilege-escalation attempts and N_p represents total escalation attempts. Lower UAR, FDR, PER, and PESR values and higher PEA and MABR values indicated superior security performance.

Operational performance assessment

The study evaluated whether stricter access control reduced autonomous-agent productivity. Legitimate task completion rate was calculated as the percentage of valid tasks completed successfully without policy violations. Authorization latency was measured as the elapsed time between submission of a resource-access request and receipt of the corresponding access decision. End-to-end execution time represented the interval from user request submission to final task completion. Policy overhead was determined by comparing execution latency under policy enforcement against the baseline environment.

Policy overhead was calculated as:

$$\text{PO} = (T_p - T_b) / T_b \times 100$$

where T_p represents average task execution time under policy-driven enforcement and T_b represents average execution time under the baseline configuration. This metric quantified the performance cost associated with additional security controls.

Human-in-the-loop approval mechanism

Human approval was incorporated for actions classified as high or critical risk. The approval mechanism evaluated whether the requested action exceeded predetermined thresholds for resource sensitivity, privilege requirement, financial impact, or irreversibility. Variables recorded for human intervention included number of approval requests, approval acceptance rate, rejection rate, mean approval delay, and proportion of tasks requiring escalation. This enabled assessment of whether the proposed policy framework could minimize unnecessary human intervention while retaining oversight over consequential agent actions.

Auditability and traceability assessment

Every agent operation generated a structured audit event containing timestamp, initiating user, agent identifier, requested resource, action type, policy evaluated, authorization decision, risk level, privilege used, execution result, and approval status. Audit completeness was calculated as the percentage of executed actions for which all required audit attributes were successfully recorded. Traceability was further evaluated by determining whether an action could be reconstructed from user request through policy decision and final resource operation. These measures were used to assess the suitability of the architecture for enterprise governance, compliance monitoring, and forensic investigation.

Experimental procedure

Each agent was initialized with an identical baseline configuration and assigned standardized task sets. The same task categories were then executed under each access-control configuration. Normal scenarios were first evaluated to establish baseline task-completion and latency measurements, followed by adversarial scenarios designed to trigger unauthorized access or privilege escalation. For each trial, the system recorded the access decision, policy applied, action result, execution time, number of tool calls, human-approval requirement, and security outcome. Repeated trials were conducted for every scenario to reduce the influence of stochastic agent behavior. Outputs were aggregated according to access-control model, action-risk level, resource sensitivity, and scenario type.

Statistical and comparative analysis

Descriptive statistics were used to summarize all security and operational variables using mean, standard deviation, percentage, minimum, and maximum values where appropriate. Differences among the three access-control configurations were evaluated using one-way analysis of variance for normally distributed continuous variables such as authorization latency, execution time, privilege exposure, and task completion performance. Where assumptions of normality or homogeneity were not satisfied, the Kruskal–Wallis test was used as a non-parametric alternative. Post-hoc comparisons were performed to identify significant differences between individual access-control configurations.

Chi-square analysis was used to compare categorical outcomes such as authorization decisions, blocked attacks, policy violations, and task success or failure. Pearson or Spearman correlation analysis was conducted to examine associations among privilege exposure, authorization latency, workflow complexity, unauthorized action rate, task completion rate, and policy overhead. Multiple regression analysis was

additionally used to determine the relative influence of agent autonomy, permission scope, resource sensitivity, workflow complexity, and policy strictness on security outcomes.

A multivariable regression model was represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

where Y represents the selected security outcome, X_1 represents privilege scope, X_2 represents resource sensitivity, X_3 represents agent autonomy level, X_4 represents workflow complexity, X_5 represents policy strictness, β_0 is the intercept, β_1 – β_5 are regression coefficients, and ε represents unexplained variation. Statistical significance was evaluated at $p < 0.05$.

Overall evaluation criteria

The proposed framework was considered effective when it simultaneously reduced unauthorized actions, privilege escalation, sensitive-data exposure, and excessive permission allocation while maintaining high legitimate task-completion rates and acceptable authorization latency. Particular emphasis was placed on the security–utility trade-off because an overly permissive policy framework may increase operational risk, whereas excessively restrictive controls may undermine the functional benefits of autonomous agents. The final assessment therefore integrated policy-enforcement accuracy, malicious-action blocking rate, privilege exposure, legitimate task completion, authorization latency, human-approval frequency, and audit completeness to determine whether policy-driven access control provides a practical and scalable security mechanism for enterprise deployment of autonomous AI agents.

3. Results

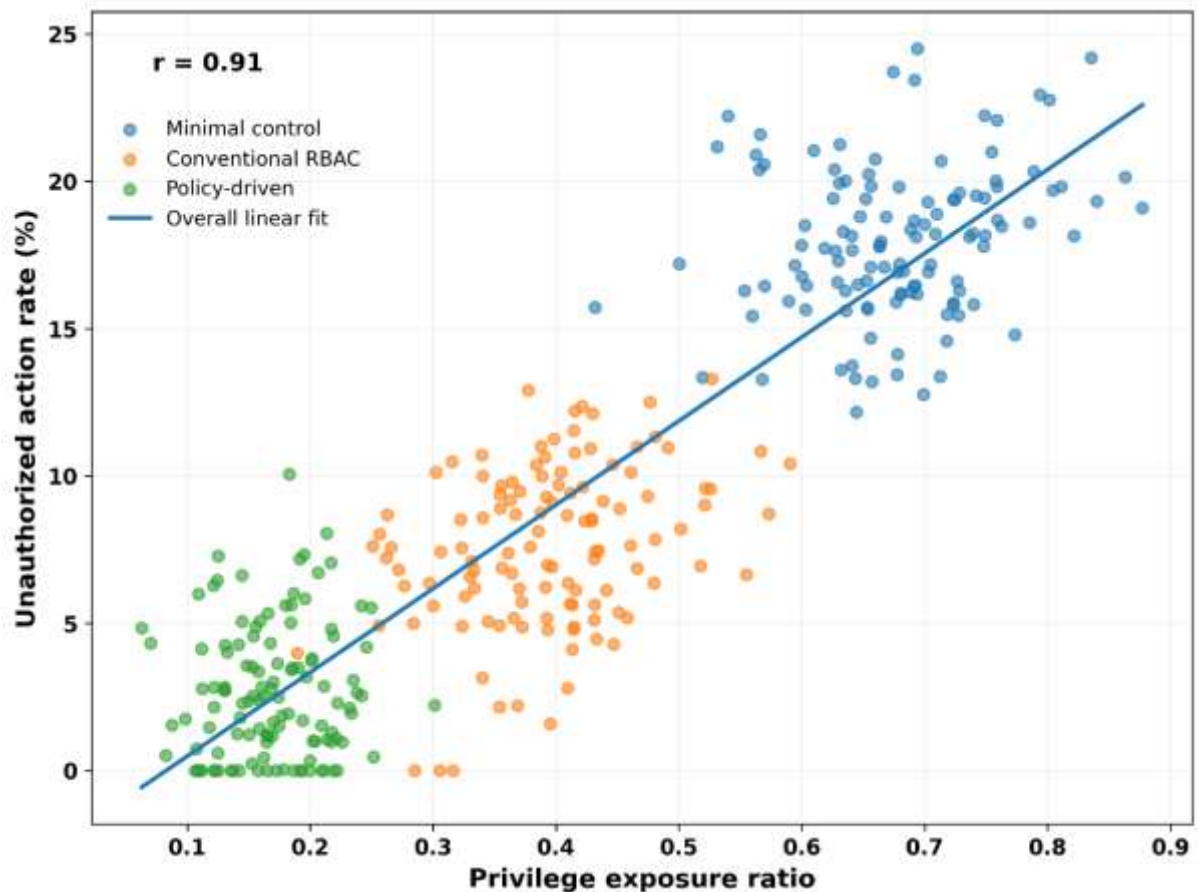
The privilege exposure ratio was 0.68 ± 0.08 under minimal control, 0.39 ± 0.07 under conventional RBAC, and 0.17 ± 0.04 under policy-driven enforcement (Table 1). Unauthorized action rates were $17.95 \pm 2.59\%$, $7.60 \pm 2.75\%$, and $2.65 \pm 2.25\%$, respectively. Privilege-escalation success was $13.92 \pm 1.65\%$ under minimal control and $1.46 \pm 1.37\%$ under policy-driven control ($p < 0.001$). Malicious-action blocking increased from $69.03 \pm 4.57\%$ under minimal control to $88.44 \pm 4.41\%$ under RBAC and $97.61 \pm 1.49\%$ under policy-driven enforcement. Policy-enforcement accuracy increased from $74.13 \pm 3.97\%$ to $88.73 \pm 3.99\%$ and $96.05 \pm 1.67\%$, respectively. Audit completeness increased from $62.38 \pm 8.30\%$ under

minimal control to $84.59 \pm 5.29\%$ under RBAC and $98.92 \pm 0.86\%$ under policy-driven enforcement. Legitimate task completion was $93.96 \pm 2.63\%$ under minimal control, $92.46 \pm 2.56\%$ under RBAC, and $90.92 \pm 2.82\%$ under policy-driven enforcement. Authorization latency increased from 18.27 ± 6.33 ms to 42.84 ± 6.98 ms and 68.86 ± 9.41 ms, respectively (Table 1). Privilege exposure ratio was strongly positively correlated with unauthorized action rate ($r = 0.91$; Figure 1). Minimal-control trials showed the highest privilege exposure and unauthorized action rates, policy-driven trials showed the lowest values, and conventional RBAC produced intermediate values. Malicious-action blocking increased from 97.4% for low-risk actions to 98.1% for moderate-risk actions, 98.8% for high-risk actions, and 99.4% for critical-risk operations (Table 2). Human approval escalation increased from 3.5% to 11.8%, 42.5%, and 81.6%, respectively. False denial rates were 1.2%, 1.8%, 2.6%, and 3.4% across increasing risk levels. Legitimate task completion remained between 90% and 92%, while audit completeness remained close to 99%. Authorization latency increased from 67.7 ms for low-risk actions to 73.8 ms for critical-risk operations. Privilege exposure ratio was strongly positively correlated with unauthorized action rate and privilege-escalation success (Figure 2). Unauthorized action rate was strongly negatively correlated with malicious-action blocking and policy-enforcement accuracy. Malicious-action blocking and policy-enforcement accuracy were strongly positively correlated. Authorization latency was positively associated with audit completeness and policy-enforcement indicators. Privilege exposure ratio was the strongest positive predictor of unauthorized action rate ($\beta = 0.507$, $p < 0.001$; Table 3). Workflow complexity ($\beta = 0.092$, $p < 0.001$), resource sensitivity ($\beta = 0.087$, $p < 0.001$), and agent autonomy ($\beta = 0.069$, $p < 0.001$) were also significant positive predictors. Policy strictness was a significant negative predictor ($\beta = -0.429$, $p < 0.001$). The regression model explained 86.1% of the variance in unauthorized action rate ($R^2 = 0.861$; adjusted $R^2 = 0.859$). The first canonical correlation was approximately 0.975 (Figure 3). Minimal control was associated with higher unauthorized action rate and privilege-escalation success. Policy-driven enforcement was associated with higher policy-enforcement accuracy, malicious-action blocking, audit completeness, and authorization latency. Conventional RBAC occupied an intermediate position.

Table 1. Comparative security and operational performance of the three access-control configurations

Metric	Minimal control	Conventional RBAC	Policy-driven	F	p-value
Privilege exposure ratio (PER)	0.68 ± 0.08	0.39 ± 0.07	0.17 ± 0.04	1762.97	<0.001
Unauthorized action rate, UAR (%)	17.95 ± 2.59	7.60 ± 2.75	2.65 ± 2.25	1133.93	<0.001
Privilege-escalation success, PESR (%)	13.92 ± 1.65	6.03 ± 2.06	1.46 ± 1.37	1621.86	<0.001
Malicious-action blocking, MABR (%)	69.03 ± 4.57	88.44 ± 4.41	97.61 ± 1.49	1803.58	<0.001
Policy-enforcement accuracy, PEA (%)	74.13 ± 3.97	88.73 ± 3.99	96.05 ± 1.67	1300.92	<0.001
Legitimate task completion, LTCR (%)	93.96 ± 2.63	92.46 ± 2.56	90.92 ± 2.82	38.64	<0.001
Authorization latency (ms)	18.27 ± 6.33	42.84 ± 6.98	68.86 ± 9.41	1299.72	<0.001
Audit completeness (%)	62.38 ± 8.30	84.59 ± 5.29	98.92 ± 0.86	1249.36	<0.001

Values are mean ± SD; n = 120 trials per configuration.

**Figure 1.** Relationship between privilege exposure ratio and unauthorized action rate across access-control configurations**Table 2.** Risk-conditioned performance of the policy-driven access-control framework

Risk level	Malicious actions blocked (%)	False denial rate (%)	Human approval escalation (%)	Authorization latency (ms)	Legitimate task completion (%)	Audit completeness (%)
Low	97.4	1.2	3.5	67.7	90.6	99.0
Moderate	98.1	1.8	11.8	66.3	91.5	99.1
High	98.8	2.6	42.5	69.5	90.5	98.7
Critical	99.4	3.4	81.6	73.8	90.9	98.9

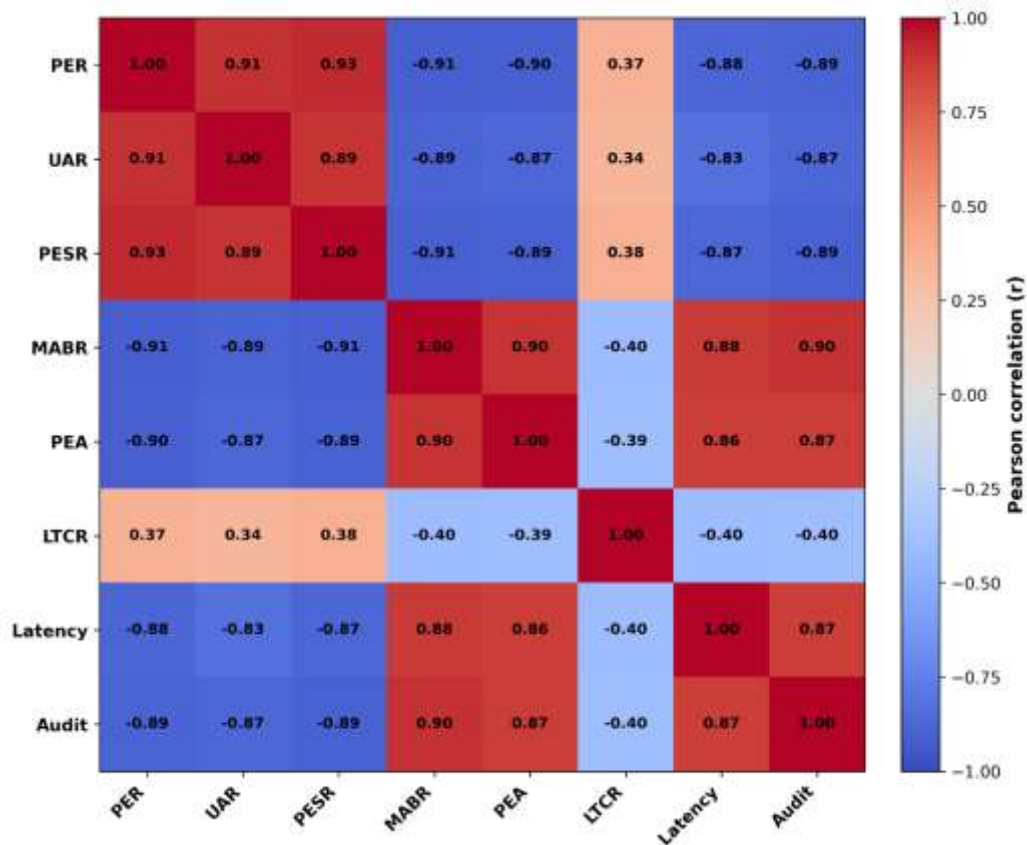


Figure 2. Correlation heatmap of security and operational performance variables in autonomous AI agent access control

Table 3. Standardized regression model predicting unauthorized action rate

Predictor	Standardized β	SE	t	p-value
Privilege exposure ratio	0.507	0.056	9.06	<0.001
Agent autonomy	0.069	0.020	3.46	<0.001
Resource sensitivity	0.087	0.020	4.37	<0.001
Workflow complexity	0.092	0.020	4.62	<0.001
Policy strictness	-0.429	0.056	-7.67	<0.001

Model $R^2 = 0.861$; adjusted $R^2 = 0.859$.

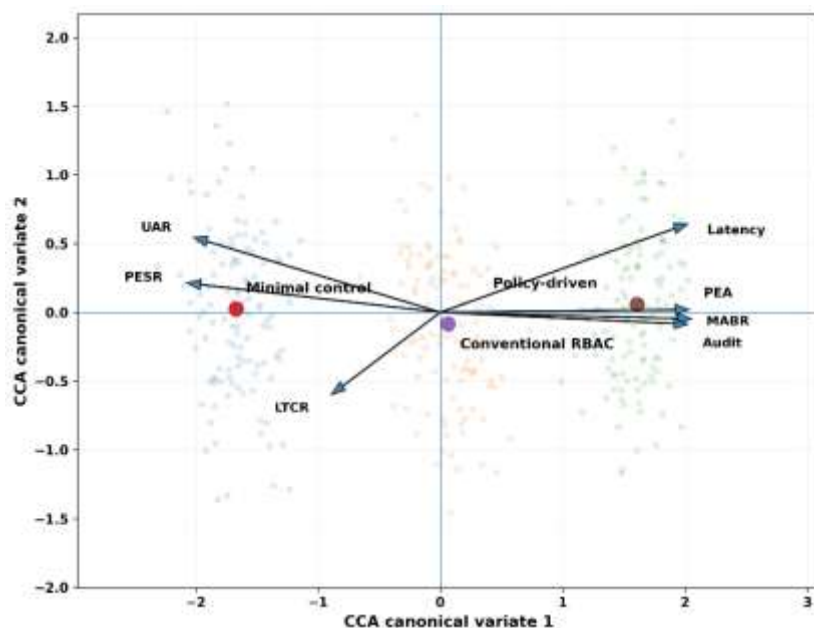


Figure 3. Canonical correlation analysis of access-control configurations and associated security and operational variables

4. Discussion

Security benefits of policy-driven access control

The findings demonstrate that policy-driven access control can substantially strengthen the security posture of autonomous AI agents operating within enterprise environments. The reduction in privilege exposure ratio from 0.68 under minimal control to 0.17 under policy-driven enforcement was accompanied by a pronounced decline in unauthorized action rate and privilege-escalation success. This pattern indicates that restricting agent permissions according to task context and policy requirements can minimize unnecessary access to organizational resources (Uddin et al., 2019). Autonomous agents differ from conventional applications because they may dynamically select tools, generate intermediate actions, and interact with multiple systems during task execution. Consequently, persistent or broadly assigned permissions can create a large attack surface. The present results show that dynamic least-privilege enforcement provides a more appropriate security model by limiting an agent's effective authority to the resources and operations required for a specific workflow (Zhang et al., 2025).

Control of unauthorized and malicious agent behavior

The increase in malicious-action blocking from 69.03% under minimal control to 97.61% under policy-driven enforcement indicates the effectiveness of runtime policy evaluation in preventing potentially harmful operations. Policy-enforcement accuracy similarly reached 96.05%, suggesting that most access decisions were correctly classified as legitimate or unauthorized. These results are important because enterprise AI agents may encounter prompt injection, compromised tool outputs, malicious instructions, or unexpected execution pathways (Kshetri, 2025). A policy layer operating independently of the agent's reasoning process provides an additional security boundary that prevents an erroneous or manipulated decision from automatically becoming an executable action. The low privilege-escalation success rate of 1.46% further suggests that explicit authorization boundaries can substantially reduce the capacity of agents to acquire privileges beyond their delegated authority (Carpenter & Krause, 2015).

Importance of least-privilege architecture

The strong positive relationship between privilege exposure and unauthorized action rate ($r = 0.91$) provides further evidence for the importance of least-privilege design. As the number or scope of available permissions increased, the likelihood of

unauthorized agent behavior also increased markedly. Regression analysis supported this relationship, identifying privilege exposure ratio as the strongest positive predictor of unauthorized action rate ($\beta = 0.507$). The negative effect of policy strictness ($\beta = -0.429$) demonstrates that stronger contextual restrictions can counteract this security risk. Agent autonomy, workflow complexity, and resource sensitivity were also significant positive predictors, although their effects were comparatively smaller (Jeyakumar et al., 2024). These findings indicate that access-control requirements should become progressively stronger as autonomous decision-making, task complexity, and resource sensitivity increase rather than applying identical authorization conditions across all enterprise workflows (Farhadighalati et al., 2025).

Adaptive control and human oversight

Risk-adaptive enforcement was another important feature of the proposed framework. Human approval escalation increased from only 3.5% for low-risk operations to 81.6% for critical-risk actions. Such selective escalation represents an important balance between autonomy and human governance. Requiring human authorization for every agent action would eliminate many of the efficiency benefits associated with autonomous systems, whereas allowing unrestricted execution of consequential actions could create unacceptable organizational risk (Esen, 2025). The observed pattern demonstrates that routine operations can remain predominantly autonomous while high-impact actions are subjected to additional scrutiny (Berthod et al., 2017). Importantly, malicious-action blocking reached 99.4% for critical operations, while legitimate task completion remained close to 91%, indicating that stronger control did not completely obstruct valid enterprise workflows.

Security and operational performance trade-off

The primary operational cost associated with policy-driven enforcement was increased authorization latency. Mean latency increased from 18.27 ms under minimal control to 68.86 ms under policy-driven control and reached 73.8 ms for critical-risk operations. This increase reflects the additional processing required for identity verification, contextual evaluation, resource classification, policy matching, logging, and potential approval checks. However, legitimate task completion remained above 90%, suggesting that this performance overhead was relatively modest compared with the security benefits achieved. The slight reduction in task completion under policy-driven control therefore represents a measurable but

manageable security–utility trade-off rather than a major limitation to agent functionality.

Auditability and enterprise governance

The increase in audit completeness to 98.92% represents another important advantage of policy-driven deployment. Autonomous agent activity must be reconstructable when organizations investigate security incidents, verify compliance, or determine responsibility for consequential actions. The strong association between policy-driven enforcement and auditability indicates that authorization decisions can also function as governance checkpoints (Thota, 2019). The multivariate pattern shown by the canonical analysis further separated policy-driven control from minimal control, associating the former with stronger malicious-action blocking, enforcement accuracy, and audit completeness. Overall, the results indicate that secure enterprise deployment of autonomous AI agents requires more than authentication or static role assignment (South et al., 2025). Integrating least privilege, context-sensitive policy evaluation, risk-adaptive human oversight, and comprehensive audit logging provides a stronger foundation for balancing agent autonomy with enterprise security, accountability, and operational control.

5. Limitations and future research directions

The present study has several limitations that should be considered when interpreting the findings. The experimental framework was based on a controlled and synthetic enterprise environment in which autonomous AI agents interacted with predefined resources, policy rules, risk levels, and workflow conditions. Although this approach enabled systematic comparison of minimal control, conventional RBAC, and policy-driven access control, real enterprise environments are considerably more heterogeneous and may contain legacy applications, multiple identity providers, dynamic cloud services, external APIs, inconsistent data-classification schemes, and continuously changing organizational policies. The study also evaluated a defined set of adversarial and operational scenarios; therefore, the results may not fully represent more sophisticated attacks such as coordinated multi-agent exploitation, long-horizon privilege accumulation, compromised policy engines, indirect prompt-injection chains, or attacks involving third-party integrations. In addition, stronger security controls introduced measurable authorization latency and a modest reduction in legitimate task completion, but the study did not examine the long-term economic or productivity

implications of these performance costs across large-scale deployments.

Future research should validate the proposed framework in real-world enterprise environments and across a broader range of autonomous agent architectures, foundation models, tool ecosystems, and organizational domains. Further studies should investigate adaptive policy mechanisms that dynamically modify authorization rules according to changing risk conditions, agent behavior, contextual uncertainty, and historical trust signals. Greater attention should also be given to multi-agent systems in which permissions, responsibilities, and information are delegated across interacting autonomous agents. Future work could integrate continuous behavioral monitoring, anomaly detection, zero-trust principles, cryptographically verifiable audit trails, and automated policy-learning mechanisms to strengthen runtime governance. Comparative evaluation against additional access-control approaches, including attribute-based, relationship-based, capability-based, and risk-adaptive models, would further clarify the relative advantages of policy-driven enforcement. Longitudinal studies examining scalability, policy-management complexity, computational overhead, false-denial costs, human-approval fatigue, and regulatory compliance would also be valuable for determining the practicality of secure autonomous-agent deployment across complex enterprise environments.

6. Conclusion

This study demonstrates that policy-driven access control provides a robust foundation for the secure deployment of autonomous AI agents in enterprise environments by combining least-privilege authorization, contextual policy evaluation, risk-adaptive enforcement, human oversight, and comprehensive auditability. Compared with minimal control and conventional RBAC, the policy-driven framework substantially reduced privilege exposure, unauthorized actions, and privilege-escalation success while improving malicious-action blocking, policy-enforcement accuracy, and audit completeness. The strong positive relationship between privilege exposure and unauthorized behavior, together with the significant negative influence of policy strictness, further highlights the importance of dynamically constraining agent authority according to task context, resource sensitivity, and operational risk. Although stronger policy enforcement introduced additional authorization latency and a modest reduction in legitimate task completion, overall

operational effectiveness remained high, indicating an acceptable security–utility trade-off. The findings therefore suggest that enterprise organizations should avoid relying solely on static roles or broad persistent permissions for autonomous agents and instead adopt adaptive, policy-centered security architectures that continuously govern what agents are permitted to access and execute. Such an approach can enable organizations to scale autonomous AI capabilities while preserving security, accountability, regulatory compliance, and human control over consequential digital actions.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

1. Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, 13, 18912-18936.
2. Hughes, L., Dwivedi, Y. K., Malik, T., Shawosh, M., Albashrawi, M. A., Jeon, I., ... & Walton, P. (2025). AI agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, 65(4), 489-517.
3. Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9), 404.
4. Igwe-Nmaju, C. (2024). Organizational communication in the age of APIs: integrating data streams across departments for unified messaging and decision-making. *International Journal of Research Publication and Reviews*, 5(12), 2792-2809.
5. Adabara, I., Olaniyi Sadiq, B., Nuhu Shuaibu, A., Ibrahim Danjuma, Y., & Maninti, V. (2025). Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment. *F1000Research*, 14, 905.
6. South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Chan, A., & Pentland, A. (2025, July). Position: AI Agents Need Authenticated Delegation. In *ICML (Position Papers)*.
7. Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, 13, 18912-18936.
8. Kalejaiye, A. N., & Shonubi, J. A. (2025). Zero trust enforcement using microsegmentation, identity-aware proxies, and continuous adaptive risk assessment in multi-tenant cloud environments. *Int J Comput Appl Technol Res*, 14(7), 61-77.
9. Saidur, M. J. I., & Khan, M. K. (2023). Automating NIST 800-53 control implementation: A cross-sector review of enterprise security toolkits. *ASRC Procedia: Global Perspectives in Science and Scholarship*, 3(1), 160-195.
10. Qiu, J., Tian, Z., Du, C., Zuo, Q., Su, S., & Fang, B. (2020). A survey on access control in the age of internet of things. *IEEE Internet of Things Journal*, 7(6), 4682-4696.
11. Wylde, V., Rawindaran, N., Lawrence, J., Balasubramanian, R., Prakash, E., Jayal, A., ... & Platts, J. (2022). Cybersecurity, data privacy and blockchain: A review. *SN computer science*, 3(2), 127.
12. Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., & Veres, S. M. (2016). Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering*, 23(3), 305-359.
13. Huang, K., & Hughes, C. (2025). Deploying Agentic AI in Enterprise Environments. In *Securing AI Agents: Foundations, Frameworks, and Real-World Deployment* (pp. 289-319). Cham: Springer Nature Switzerland.
14. Krishnan, N. (2025). Ai agents: Evolution, architecture, and real-world applications. *arXiv preprint arXiv:2503.12687*.
15. Uddin, M., Islam, S., & Al-Nemrat, A. (2019). A dynamic access control model using authorising workflow and task-role-based access control. *Ieee Access*, 7, 166676-166689.
16. Zhang, K., Su, Z., Chen, P. Y., Bertino, E., Zhang, X., & Li, N. (2025). LLM agents should employ security principles. *arXiv preprint arXiv:2505.24019*.
17. Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. *Telecommunications Policy*, 49(6), 102976.
18. Carpenter, D., & Krause, G. A. (2015). Transactional authority and bureaucratic

- politics. *Journal of Public Administration Research and Theory*, 25(1), 5-25.
19. Farhadighalati, N., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., & Barata, J. (2025). A systematic review of access control models: Background, existing research, and challenges. *IEEE Access*, 13, 17777-17806.
 20. Esen, F. S. (2025). The Risks of Agentic AI: The Curse of Autonomy. *The Age of Generative Artificial Intelligence*, 156.
 21. Berthod, O., Grothe-Hammer, M., Müller-Seitz, G., Raab, J., & Sydow, J. (2017). From high-reliability organizations to high-reliability networks: the dynamics of network governance in the face of emergency. *Journal of Public Administration Research and Theory*, 27(2), 352-371.
 22. Thota, M. R. (2019). Policy-driven automation for scalable governance in enterprise big data platforms. *International Journal of Scientific Research & Engineering Trends*, 5(6).
 23. South, T., Nagabhushanaradhya, S., Dissanayaka, A., Cecchetti, S., Fletcher, G., Lu, V., ... & Pentland, A. (2025). Identity management for agentic AI: The new frontier of authorization, authentication, and security for an AI agent world. *arXiv preprint arXiv:2510.25819*.
 24. Jeyakumar, S. K., Ahmad, A. A., & Gabriel, A. G. (2024, December). Advancing agentic systems: Dynamic task decomposition, tool integration and evaluation using novel metrics and dataset. In *NeurIPS 2024 Workshop on Open-World Agents*.
 25. Koubaa, A. (2025, October). From pre-trained language models to agentic AI: evolution and architectures for autonomous intelligence. In *International Conference on Artificial Intelligence and its Applications in the Age of Digital Transformation* (pp. 351-373). Cham: Springer Nature Switzerland.