

Zero Trust Agency: Mitigating the Confused Deputy in Autonomous AI Systems

Rajkumar Vaghashiya
Saarland University
AI Safety Saarland

Anshul
Saarland University
AI Safety Saarland

With
Apart Research

Abstract

As AI agents transition from passive chatbots to autonomous actors capable of executing code, they introduce "Confused Deputy" vulnerabilities where agents operating with broad privileges are manipulated via prompt injection. We propose a Zero Trust Architecture centered on Identity Propagation and Least Privilege Tooling. Our MVP, simulating different user roles, successfully blocked 100% of privilege escalation attempts by enforcing the "On-Behalf-Of" principle, where agents cryptographically bind actions to the human user's identity. This architecture effectively closes the governance gap in agentic workflows.

Keywords: AI security, defensive infrastructure, Identity Propagation, Least Privilege Tooling

1. Introduction

The emergence of agentic AI creates a critical security vacuum. Unlike chatbots, agents execute code and modify data. However, this autonomy introduces severe risks; reports indicate that **97% of AI-related breaches** occur in environments lacking granular access controls [1]. Current implementations often grant agents static, high-level permissions, allowing attackers to exploit the "**Confused Deputy**" vulnerability—tricking the agent into abusing its authority [3]. Motivated by threats like the GTG-1002 campaign [2], our architecture implements a defense where agents possess no standing privileges and must "borrow" identity from the user.

To address this, we use the "**On-Behalf-Of**" (**OBO**) principle which dictates that the agent acts as a transparent proxy with **zero standing privileges**. It must cryptographically "borrow"

the user's identity for each specific tool invocation. This ensures the agent's capabilities are mathematically constrained by the user's verified permissions, rendering various "jailbreaks" ineffective against unauthorized resources[3].

2. Methods

We engineered a secure agent ecosystem compliant with the NIST AI Risk Management Framework[5].

- **Architecture: Identity Propagation** using **RFC 8693 (OAuth 2.0 Token Exchange)**. The Agent Orchestrator receives a user's token and exchanges it for a downstream tool token, maintaining a cryptographic **Audit Trail**.
- **Simulated User Roles:**
 - **Admin:** Full access (github:*).
 - **Developer:** Read/write (repo:read, issue:create), NO delete.
 - **Finance/Marketing:** Read-only access to public repositories.
- **Attack Simulations:** We tested vectors including Prompt Injection ("Ignore instructions"), Privilege Escalation, and Data Exfiltration against these roles[4].

3. Results

The Zero Trust architecture deterministically neutralized attacks by enforcing scope boundaries at the infrastructure layer.

- **Blocked Privilege Escalation:** A Finance user attempting to access internal source code was denied.
 - *Log:* DENIED: Missing scope: github:read_private_repos | User: finance@company.com
- **Authorized Execution:** A Developer user successfully created an issue.
 - *Log:* ALLOWED: create_issue | User: developer@company.com | Chain: developer -> ai_agent -> mcp_github
- **Prompt Injection Mitigation:** Direct injection attempts by a Developer to "delete all repos" failed because the propagated token lacked the delete scope, regardless of the prompt's persuasiveness.

The simulation results can be viewed [here](#).

4. Discussion and Conclusion

Our findings confirm that **architectural identity controls** are superior to probabilistic alignment techniques. By decoupling the agent's intelligence from its authority via the OBO principle, we significantly mitigate the "Confused Deputy" risk. The agent cannot be tricked into doing what it is not authorized to do, shifting the defense mechanism from "monitoring" to "architecture".

5. References

1. Cost of a Data Breach Report 2025, IBM Security, accessed November 23, 2025, <https://www.ibm.com/reports/data-breach>
2. Disrupting the first reported AI-orchestrated cyber espionage campaign - Anthropic, accessed November 23, 2025, <https://www.anthropic.com/news/disrupting-AI-espionage>
3. 3 Takeaways from the OWASP Agentic AI Security Research - Entro, accessed November 23, 2025, <https://entro.security/blog/agentic-ai-owasp-research/>
4. Security for AI Agents: Protecting Intelligent Systems in 2025, Obsidian Security Team, accessed November 23, 2025, <https://www.obsidiansecurity.com/blog/security-for-ai-agents>
5. NIST. (2025). AI Risk Management Framework.

6. Appendix

Security Considerations

- **Latency:** The OBO token exchange introduces network overhead (~100ms); **Token Caching** strategies are required for real-time performance.
- **Granularity:** Security is limited by the underlying Identity Provider; if a user is a Global Admin, the agent inherits those risks. **Just-in-Time (JIT)** elevation is suggested for future improvements.

- **Supply Chain:** While OBO protects access, it does not prevent "Slopsquatting" (malicious package installation). This requires separate **Tool Allow-listing**.