Security Risk in Autonomous AI Agents: A Survey of Threats, Attack Surfaces, and Defenses

Cheng Wen[1], Hua Zheng[2], Yuhang Guo[1], Muhammad Sadiq[3], Kuanishbay Sadatdiynov[4], Jawwad Ahmed Shamsi[5], and Shengchao Qin[1]

[1]*Xidian University,* [2]*Guangzhou University of Software,* [3]*Shenzhen University of Information Technology,* [4]*Nukus State Technical University,* [5]*National University of Computer and Emerging Sciences Karachi*

Introduction

1. Introduction

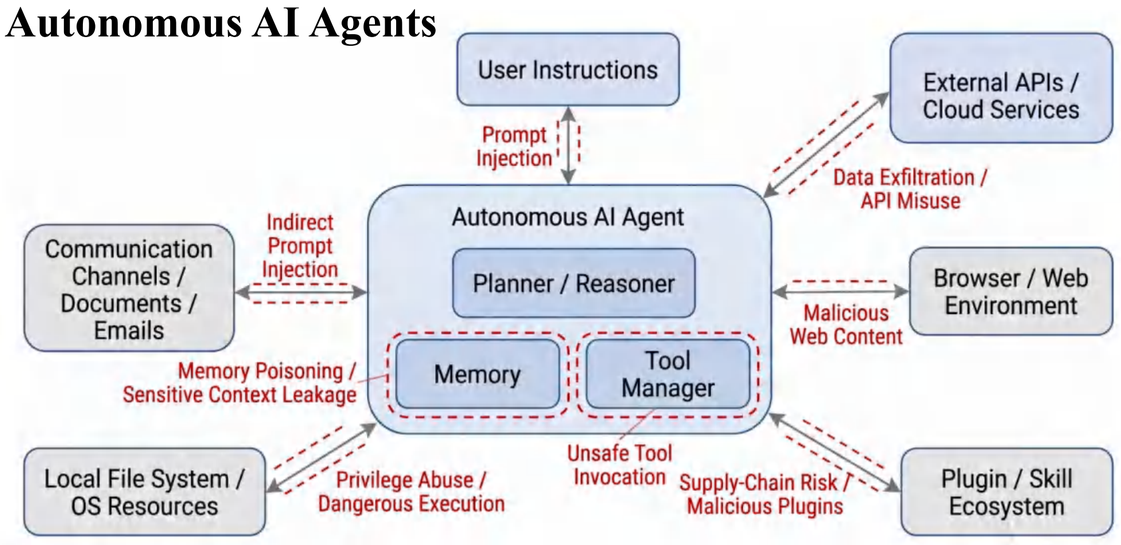
Autonomous AI agents have evolved from conversational assistants into action-oriented systems capable of planning, invoking tools, accessing local and remote resources, and executing complex workflows with limited human intervention.

However, increased autonomy introduces a broader attack surface beyond traditional software and standalone LLM applications. Agent vulnerabilities may lead to real-world harmful consequences, including unauthorized file access, credential leakage, malicious command execution, and persistent system compromise.

Therefore, securing autonomous AI agents requires a system-level perspective that considers not only model reasoning, but also tool usage, memory, data access, external interaction, plugin ecosystems, and deployment configurations.

High-Level Architecture

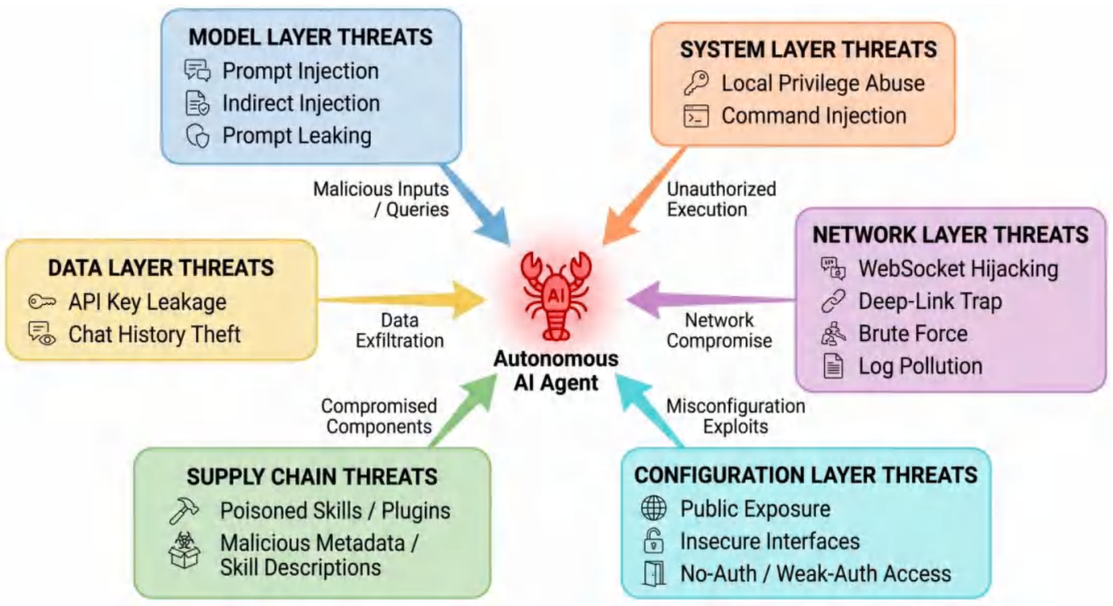
2. High-Level Architecture and Attack Surface of Autonomous AI Agents



Taxonomy

3. Threats and Vulnerabilities in Autonomous AI Agents

To keep this threat landscape clear, we organize representative risks into six layers: model, system, data, network, supply chain, and configuration.



Attack Surfaces

4. Attack Surfaces and Security Implications

Autonomous AI agents introduce interconnected attack surfaces across:

- Input: Malicious prompts and external content can manipulate reasoning.
- Execution: Tool misuse may lead to unsafe commands, file access, or data leakage.
- Ecosystem: Plugins and APIs create supply-chain and trust risks.
- Impact: Agent autonomy, persistence, and privilege concentration amplify security consequences.

Table 3: Defense strategies for autonomous AI agents.

Category	Main idea	Protects against	Limitation
Input & reasoning	Filter and separate unsafe inputs	Prompt injection, unsafe reasoning	Hard to catch subtle attacks
Execution control	Restrict permissions and isolate actions	Dangerous commands, file misuse	Reduces autonomy
Runtime monitoring	Detect abnormal behavior during execution	Hidden misuse, cross-layer attacks	May cause overhead
Supply chain	Review plugins and control dependencies	Malicious plugins, update abuse	Depends on review quality
Deployment security	Use secure defaults and strong authentication	Public exposure, weak access control	Misconfiguration is still common

Defense Strategies

5. Defense Strategies

Defending autonomous AI agents requires more than filtering unsafe outputs. Because agents combine reasoning, execution, memory, and external interaction, effective protection must cover multiple stages of the agent lifecycle.

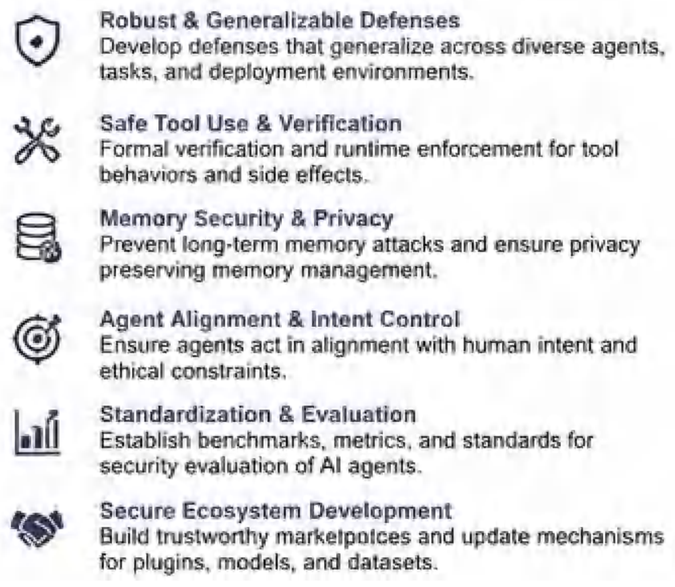
We summarize existing defense strategies from five perspectives: input and reasoning, execution and access control, monitoring and runtime protection, ecosystem and supply-chain security, and secure configuration and deployment.



Open Challenges

6. Open Challenges and Future Directions

Despite growing attention to agent security, many challenges remain open. First, there is a fundamental tension between autonomy and control. Autonomous AI agents are most useful when they can act flexibly and efficiently, but stronger security mechanisms often reduce this flexibility by adding restrictions, confirma_x0002_tions, or monitoring overhead.



Conclusion. Autonomous AI agents bring tremendous potential but also unprecedented security risks due to their autonomy, tool use, memory, and system access. A comprehensive understanding of threats, attack surfaces, and defense strategies is essential for building safer, more controllable, and more trustworthy agentic systems.

The 23rd International Conference on [illegible] Information Systems and Applications, Kunming, [illegible], 2026

Paper ID: P00439 **Contact email: wencheng@xidian.edu.cn** **Phone: 19147951419**