January 8, 2026

ZombieAgent: The Agentic Revolution Comes with Malicious Gifts

> **Key Insights:**
>
> - Agentic AI creates a powerful new attack surface beyond traditional security controls
> - Zero-click indirect prompt injection enables invisible compromise and data exfiltration
> - Agent memory manipulation turns AI agents into persistent insider threats
> - Compromised agents can self-propagate across organizations and ecosystems
> - Conventional enterprise defenses and LLM guardrails are not enough to detect and contain persistent, service-side, indirect prompt injection attacks such as ZombieAgent

Executive Summary

Organizations across industries are rapidly adopting autonomous AI agents that can read emails, interact with corporate systems, trigger workflows and make decisions without human intervention. This shift marks the beginning of what many are calling the "agentic economy." These agents promise major productivity benefits, yet they also introduce a new and largely unregulated attack surface. Recent incidents, such as Radware's ShadowLeak vulnerability (1, 2), underscore how easily attackers can exploit this emerging ecosystem. By embedding hidden instructions into emails, documents, or other untrusted data, adversaries can manipulate an AI agent's behavior without user interaction and exfiltrate an organization's sensitive data directly from the AI provider's cloud infrastructure. These attacks leave no traces on the user's device, bypass network monitoring and evade traditional detection systems.

The risks extend beyond data leakage. Newly uncovered advanced forms of indirect prompt injection (IPI) detailed by Radware's Security Researchers in their ZombieAgent report allow attackers not only to extract information but to implant persistent logic into an agent's long-term memory, effectively taking over the agent and turning it into a silent insider operating within the enterprise. Because these agents interact with corporate environments and often have access to sensitive data, the impact can be significant. The combination of autonomy, connectivity and limited visibility makes agentic AI an attractive target for cybercriminals and a potential vector for long-term compromise.

The Threat

The threat landscape surrounding agentic AI is expanding quickly and is defined by attackers' ability to manipulate the agent's perception of what constitutes a legitimate instruction. Indirect prompt injection, the core vulnerability class in this ecosystem, exploits the fact that large language models cannot inherently distinguish between trusted instructions and untrusted user content. If an email, document, or webpage contains hidden directives, such as instructions concealed via formatting tricks, invisible HTML, or encoded payloads, the agent may interpret them as valid commands. When the agent later processes that content while performing a routine task, it may unknowingly carry out an attacker's instructions.

A critical complication arises from the fact that many agent actions are executed directly from the AI provider's cloud infrastructure rather than the user's device. This was demonstrated clearly in the ShadowLeak attack, where the exfiltration originated entirely from OpenAI's cloud systems. Because the actions were service-side, none of the usual enterprise security controls were engaged. No logs appeared on the endpoint, no suspicious traffic was recorded on the corporate network, and no traditional data leak detection (DLP) solution had visibility into the agent's activity. This architecture is now common across multiple AI platforms, meaning the same invisibility applies to other providers as well.

In more advanced attacks, the scope extends beyond a single exfiltration event. If an attacker manages to influence how an agent stores information in its memory or working environment, the compromise may become persistent. Agents designed to retain context across sessions or optimize themselves by remembering user preferences can be tricked into storing attacker-defined rules. Once implanted, these rules may cause the agent to execute malicious operations every time it is invoked, often before responding to legitimate user queries. Because this behavior is embedded in memory, the compromise does not require the attacker to re-engage. It can continue operating indefinitely, collecting information quietly and sending it to external servers.

Propagation adds another layer of risk. A compromised agent may be instructed to harvest additional email addresses or extract contacts from a user's mailbox. It can then forward a similar malicious payload to new targets. The mechanism allows the attack to spread organically within an organization and its partner ecosystem, much like traditional email-borne malware but with greater automation and accuracy. This turns a single compromised agent into an entry point for a wider campaign.

Zero-Click Indirect Prompt Injection Attack Flow

A typical attack begins with the delivery of a crafted email or document that is indistinguishable from harmless communication. Hidden within the content are embedded instructions that instruct the agent to perform actions such as collecting inbox data, reading sensitive files, or contacting external endpoints. At no point does the user need to click anything or interact with the message. The mere act of asking the agent to summarize the inbox or search for relevant communications triggers the exploit.

Once the malicious content is processed, the agent begins to execute the embedded instructions. Because it operates in the provider's cloud environment, the outbound connections originate from trusted infrastructure. Firewalls, secure web gateways, proxies and endpoint protection tools see no evidence of wrongdoing. To the victim, the interaction appears normal. The agent responds with a standard summary or output, but in the background, it has already completed the attacker's secondary tasks.

If the payload includes memory-modification logic, the agent may rewrite its internal rules or notes, which creates recurring malicious activity. From that moment on, every new user request is accompanied by unauthorized actions carried out silently on behalf of the attacker. In propagation scenarios, the agent may gather additional addresses, generate new outreach messages, and automatically distribute similar payloads to additional recipients. This enables fast, quiet, worm-like expansion through corporate environments.

Why Traditional Security Controls Fail

This class of attack exposes fundamental weaknesses in enterprise security architecture. Existing tools are designed around human-triggered actions that originate from endpoints or corporate networks. They are not built to monitor what happens inside an AI service or how an agent interprets untrusted content. Even advanced security systems cannot protect against actions entirely within a cloud provider's infrastructure, because the organization has no visibility into or control over them.

Moreover, guardrails built by AI vendors are largely rule-based and reactive. They typically rely on patterns, filters and heuristics to block dangerous requests. Attackers can easily design prompts that technically comply with these rules while still achieving malicious goals. For example, ZombieAgent used a character-by-character exfiltration technique and indirect link manipulation to circumvent the guardrails OpenAI implemented to prevent its predecessor, ShadowLeak, from exfiltrating sensitive information. Because the LLM has no inherent understanding of intent and no reliable boundary between system instructions and external content, these attacker methods remain effective despite incremental vendor improvements.

The combination of missing visibility, unrestricted execution capabilities and brittle guardrails makes indirect prompt injection one of the most significant emerging risks associated with agentic AI.

Recommended Mitigation Actions

Organizations adopting agentic AI must adapt their security posture to account for the unique behavior and capabilities of autonomous agents. These agents should be treated as privileged digital identities with the potential to perform broad and sensitive operations. Limiting their access, monitoring their actions and sanitizing their inputs is essential.

A core requirement is to restrict the range and scope of data that an agent can read and the actions it can perform. Separating reading permissions from execution capabilities reduces the potential impact of a compromise. All inbound content from untrusted sources should be cleaned, normalized or converted to safe plain text before being passed to an agent. Without this sanitization, any email or document could act as an unmonitored attack surface.

Visibility must be enhanced by logging all agent actions, especially those involving data access or external requests. Behavioral monitoring, rather than simple rule-based checks, is required to detect when an agent's actions diverge from user intent. Red-teaming should be conducted before deployment, focusing specifically on zero-click IPI exploitation, memory corruption, propagation mechanisms and service-side exfiltration.

Governance is equally critical. Organizations must establish policies that define which systems agents may access, under what conditions, and with what scopes. Permissions should be reviewed regularly, and access must not be granted permanently without justification. Vendor assessments should include questions about how agentic AI is isolated, monitored and protected against prompt injection. Proper internal controls combined with informed procurement requirements can significantly reduce risk.

Conclusion

The emergence of agentic AI represents a major shift in enterprise technology. These systems bring powerful automation and operational benefits, but they also create new opportunities for attackers. Zero-click prompt injection, persistent memory manipulation and service-side exfiltration highlight the seriousness of the threat. Organizations must adapt quickly, recognizing that agents are not simple productivity tools but privileged entities capable of acting broadly and silently within corporate environments.

To realize the benefits of agentic AI while avoiding its risks, security must be built into every stage of adoption: from procurement and configuration to monitoring, mitigating and incident response. In this new era, organizations that treat AI agents as fully-fledged identities with corresponding controls, oversight and governance will be best positioned to protect their data, employees and operations.

EFFECTIVE DDOS PROTECTION ESSENTIALS

Hybrid DDoS Protection – Use on-premises and cloud DDoS protection for real-time DDoS attack prevention that also addresses high-volume attacks and protects from pipe saturation

Behavioral-Based Detection – Quickly and accurately identify and block anomalies while allowing legitimate traffic through

Real-Time Signature Creation – Promptly protect against unknown threats and zero-day attacks

Web DDOS Tsunami Protection – Automated immediate detection and mitigation of Web DDOS encrypted high RPS and morphing attacks

A Cybersecurity Emergency Response Plan – Turn to a dedicated emergency team of experts who have experience with Internet of Things security and handling IoT outbreaks

Intelligence on Active Threat Actors – High fidelity, correlated and analyzed data for preemptive protection against currently active known attackers

For further network and application protection measures, Radware urges companies to inspect and patch their network to defend against risks and threats.

EFFECTIVE WEB APPLICATION SECURITY ESSENTIALS

Full OWASP Top-10 coverage against defacements, injections, etc.

Low false positive rate using negative and positive security models for maximum accuracy

Auto-policy generation capabilities for the widest coverage with the lowest operational effort

Bot protection and device fingerprinting capabilities to overcome dynamic IP attacks and achieve improved bot detection and blocking

Securing APIs by filtering paths, understanding XML and JSON schemas for enforcement, and using activity tracking mechanisms to trace bots and guard internal resources

Flexible deployment options including on-premises, out-of-path, virtual or cloud-based

LEARN MORE AT RADWARE’S SECURITY RESEARCH CENTER

To know more about today’s attack vector landscape, understand the business impact of cyberattacks, or learn more about emerging attack types and tools, visit Radware’s Security Research Center. Additionally, visit Radware’s Quarterly DDoS & Application Threat Analysis Center for quarter-over-quarter analysis of DDoS and application attack activity based on data from Radware’s cloud security services and threat intelligence.

THIS REPORT CONTAINS ONLY PUBLICLY AVAILABLE INFORMATION, WHICH IS PROVIDED FOR GENERAL INFORMATION PURPOSES ONLY. ALL INFORMATION IS PROVIDED “AS IS” WITHOUT ANY REPRESENTATION OR WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES THAT THIS REPORT IS ERROR-FREE OR ANY IMPLIED WARRANTIES REGARDING THE ACCURACY, VALIDITY, ADEQUACY, RELIABILITY, AVAILBILITY, COMPLETENESS, FITNESS FOR ANY PARTICULAR PURPOSE OR NON-INFRINGEMENT. USE OF THIS REPORT, IN WHOLE OR IN PART, IS AT USER’S SOLE RISK. RADWARE AND/OR ANYONE ON ITS BEHALF SPECIFICALLY DISCLAIMS ANY LIABILITY IN RELATION TO THIS REPORT, INCLUDING WITHOUT LIMITATION, FOR ANY DIRECT, SPECIAL, INDIREC, INCIDENTAL, CONSEQUENTIAL, OR EXAMPLARY DAMAGES, LOSSES AND EXPENSES ARISING FROM OR IN ANY WAY RELATED TO THIS REPORT, HOWEVER CAUSED, AND WHETHER BASED ON CONTRACT, TORT (INCLUDING NEGLIGENCE) OR OTHER THEORY OF LIABILITY, EVEN IF IT WAS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES, LOSSES OR EXPENSES. **CHARTS USED OR REPRODUCED SHOULD BE CREDITED TO RADWARE**