INTELLIGENCE BRIEFING
Security Command Center
TLP:CLEAR
2026-06-16 08:12 UTC

OpenClaw AI Agent Compromised by Both Injection and Social Engineering: Two Attack Paths, One Structural Problem

SECURITY ANALYSIS | HIGH | CVSS 7.5

SCC Item ID	SCC-STY-2026-0214
Type	Security Analysis
Severity	HIGH
CVSS Base Score	7.5
Affected Products	OpenClaw AI agent (pre-2026.4.23), Google Gemini 3.1 Pro, OpenAI Codex GPT-5.4, OpenClaw Slack/Discord/Matrix/Zalo/Microsoft Teams channel extensions
Published	2026-06-11T13:46:32
Discovery Source	Rss

Executive Summary

Researchers at Imperva and Varonis independently demonstrated two separate attack paths against OpenClaw, a widely deployed self-hosted AI agent: a prompt injection flaw (patched in version 2026.4.23) and an unpatched social engineering vector that bypasses sender verification by exploiting mutable display names in channel integrations. Both paths exploit what researchers call the 'lethal trifecta' - broad permissions, unsanitized content ingestion, and outbound data transmission - enabling attacker code execution and exfiltration of credentials including AWS IAM keys and database connection strings. Neither vector has been assigned a CVE at this time. This incident signals that AI agents are now a distinct attack surface with systemic trust architecture problems that patches alone cannot resolve.

Technical Analysis

Imperva and Varonis disclosed two structurally distinct attack paths against OpenClaw, a self-hosted AI agent with integrations across Slack, Discord, Matrix, Zalo, and Microsoft Teams.

The first path exploits unsanitized message object fields to achieve prompt injection (CWE-77, mapped to MITRE T1190 and T1059). By embedding malicious instructions within serialized message content, an attacker causes the agent to execute attacker-controlled commands and exfiltrate data to an external address via the agent's own outbound channel (T1041). OpenClaw addressed this in version 2026.4.23 by patching the serialization flaw. No CVE has been assigned to this vector.

The second path, and the more structurally significant one, remains unpatched. Researchers demonstrated that OpenClaw's allowlist resolution operates against mutable display names rather than cryptographically verified sender identities (CWE-345, CWE-284). An attacker who sets their display name to match an allowlisted entity can send messages the agent treats as trusted. Paired with urgency or routine pretexts (T1566), this constitutes agent phishing: the agent acts on instructions from what it believes is an authorized sender without any ground-truth identity verification. The vulnerability maps to T1036 (Masquerading) because the trust bypass is display-name spoofing, not credential theft. This vector remains unassigned and unpublished upstream.

Both paths converge on the same consequence: the agent, holding broad access permissions, executes attacker instructions and transmits credentials, AWS IAM keys, database connection strings, and SSH credentials (T1552) to attacker-controlled destinations.

Researchers frame the underlying condition as a 'lethal trifecta': excessive permissions + unsanitized content ingestion + outbound transmission capability. This framing has direct implications beyond OpenClaw. Any AI agent architecture that combines these three properties is structurally exploitable regardless of which specific injection or social engineering technique is used. The unpatched social engineering vector is particularly significant because it operates entirely within normal communication channels, produces no anomalous network signatures at the point of instruction delivery, and degrades as a detection target over time as agents are trained to be more responsive to natural-language urgency cues.

Organizations running OpenClaw prior to 2026.4.23 on channel integrations should treat both vectors as active risks. The patch closes the injection path; it does not address the identity verification gap.

Action Checklist

1. Step 1: Assess exposure - identify all deployments of OpenClaw across your environment, including self-hosted instances connected to Slack, Discord, Matrix, Zalo, or Microsoft Teams channel extensions; confirm which version is running against the 2026.4.23 patch boundary.
2. Step 2: Apply the available patch - upgrade all OpenClaw instances to version 2026.4.23 or later to close the prompt injection vector; document which instances have been updated and which remain on affected versions, per CIS 7.3 (Perform Automated Operating System Patch Management) and CIS 7.4 (Perform Automated Application Patch Management).
3. Step 3: Audit agent permissions against least-privilege principle - review and reduce the access permissions granted to OpenClaw and any other AI agents in your environment; agents should hold only the minimum permissions required for their defined function, per NIST AC-6 (Least Privilege) and CIS 5.4 (Restrict Administrator Privileges to Dedicated Administrator Accounts); revoke standing access to credential stores, IAM configurations, and sensitive datasets.
4. Step 4: Evaluate sender-identity verification controls - determine whether your AI agent configurations resolve trust based on display names or verified sender identities; where display-name-based allowlists are in use, assess whether cryptographic sender verification or out-of-band confirmation requirements can be introduced; this addresses the unpatched social engineering vector pending an upstream fix, and maps to NIST AC-3 (Access Enforcement) and NIST AC-4 (Information Flow Enforcement).
5. Step 5: Instrument agent activity for behavioral anomalies - enable and review audit logs covering agent-initiated outbound connections, credential access events, and execution of commands or scripts; alert on agent activity that occurs outside established operational baselines, per NIST AU-2 (Event Logging), AU-6 (Audit Record Review, Analysis, and Reporting), and CIS 8.2 (Collect Audit Logs).

6. Step 6: Update threat model - register the 'lethal trifecta' pattern (broad permissions + unsanitized content ingestion + outbound transmission) as a structural risk category applicable to all AI agents in your environment, not only OpenClaw; incorporate T1036, T1059, T1041, T1552, and T1566 into your AI agent threat profile.

7. Step 7: Brief leadership - communicate that the social engineering path remains unpatched upstream and that organizational risk reduction depends on compensating controls (permission reduction, sender verification, behavioral monitoring) until OpenClaw issues a fix; frame this as a systemic AI agent trust architecture issue, not an isolated software bug. Do not assume a prompt timeline for the upstream fix.

8. Step 8: Monitor for vendor patch and researcher follow-up - assign responsibility for weekly review of OpenClaw release notes at github.com/openclaw/openclaw/releases and openclaw.com.au/updates, and monitor Imperva and Varonis for published indicators or additional technical detail. Alert security leadership upon upstream fix announcement.

Detection Guidance

Detection for the prompt injection path (patched in 2026.4.23): Review historical logs for agent activity on versions prior to 2026.4.23. Look for command or script execution events (T1059) initiated by the agent that are not traceable to a recognized user instruction in context. Flag any outbound data transfer events (T1041) from the agent process to external addresses not in your approved destination list. Cross-reference NIST AU-3 (Content of Audit Records) requirements; logs should capture what action occurred, when, by which process, and to which destination.

Detection for the social engineering path (unpatched): This vector is harder to detect at the point of instruction delivery because messages arrive through normal channels. Focus detection on the agent's response behavior rather than the inbound message. Hunt for: (1) agent-initiated access to credential stores (AWS IAM, database connection strings, SSH key files) outside normal operational workflows; (2) outbound connections from the agent process to destinations not previously observed or not in an approved allowlist; (3) agent execution of instructions framed with urgency language or unusual task types - review agent interaction logs for anomalous instruction patterns. NIST AU-6 (Audit Record Review, Analysis, and Reporting) provides the framework for this ongoing review cadence.

Credential exfiltration hunting: T1552 (Unsecured Credentials) exploitation should surface in access logs for IAM key stores, secrets managers, and database configuration files. Alert on any process, including the agent process, reading multiple credential objects in a short time window outside a scheduled or documented workflow. CIS 8.2 (Collect Audit Logs) and NIST AU-12 (Audit Record Generation) establish baseline requirements for the log sources this hunting requires.

Behavioral baseline: Establish a documented baseline of normal OpenClaw agent behavior - which channels it reads, which systems it accesses, what data it transmits outbound, and at what frequency. Deviations from this baseline are the primary detection signal for both the patched and unpatched vectors. MITRE D3FEND supplementary countermeasures applicable here include D3-LAM (Local Account Monitoring) for agent account activity and D3-UAP (User Account Permissions) review to limit the blast radius of a successful compromise.

Indicators of Compromise

Type	Value	Context	Confidence
TOOL	`Pending — refer to Imperva and Varonis research disclosures for published indicators`	Imperva and Varonis independently demonstrated both attack paths; their published research may include payload samples, injected instruction patterns, or exfiltration destination indicators not reproduced in available source summaries	LOW

Framework Mappings

MITRE-ATTACK

- **T1078** — Valid Accounts
- **T1552** — Unsecured Credentials
- **T1036** — Masquerading
- **T1059** — Command and Scripting Interpreter
- **T1566** — Phishing
- **T1041** — Exfiltration Over C2 Channel
- **T1190** — Exploit Public-Facing Application
- **T1189** — Drive-by Compromise

NIST-800-53R5

- **AC-2** — Account Management
- **AC-6** — Least Privilege
- **IA-2** — Identification and Authentication (Organizational Users)
- **IA-5** — Authenticator Management
- **CM-7** — Least Functionality
- **SI-3** — Malicious Code Protection
- **SI-4** — System Monitoring
- **SI-7** — Software, Firmware, and Information Integrity
- **AT-2** — Literacy Training and Awareness
- **CA-7** — Continuous Monitoring
- **SC-7** — Boundary Protection
- **SI-8** — Spam Protection
- **CA-8** — Penetration Testing
- **RA-5** — Vulnerability Monitoring and Scanning
- **SI-2** — Flaw Remediation
- **SI-10** — Information Input Validation
- **AC-3** — Access Enforcement
- **SC-28** — Protection of Information at Rest

OWASP-TOP10-2021

- **A08:2021** — Software and Data Integrity Failures
- **A03:2021** — Injection
- **A01:2021** — Broken Access Control

CIS-V8

- **2.5** — Allowlist Authorized Software
- **16.10** — Apply Secure Design Principles in Application Architectures
- **6.1** — Establish an Access Granting Process
- **6.2** — Establish an Access Revoking Process
- **6.3** — Require MFA for Externally-Exposed Applications
- **7.3** — Perform Automated Operating System Patch Management
- **7.4** — Perform Automated Application Patch Management
- **14.2** — Train Workforce Members to Recognize Social Engineering Attacks

SOC2-TSC

- **CC6.1** — The entity implements logical access security software, infrastructure, and architectures over protected information assets

HIPAA-SECURITY

- **164.312(a)(1)** — Access Control
- **164.312(d)** — Person or Entity Authentication
- **164.308(a)(5)(i)** — Security Awareness and Training

ISO-27001-2022

- **A.8.8** — Management of technical vulnerabilities
- **A.5.34** — Privacy and protection of personal information

MITRE ATT&CK Mapping

Technique ID	Technique Name	Tactic
T1078	Valid Accounts	Defense-Evasion
T1552	Unsecured Credentials	Credential-Access
T1036	Masquerading	Defense-Evasion
T1059	Command and Scripting Interpreter	Execution
T1566	Phishing	Initial-Access
T1041	Exfiltration Over C2 Channel	Exfiltration
T1190	Exploit Public-Facing Application	Initial-Access

Technique ID	Technique Name	Tactic
T1189	Drive-by Compromise	Initial-Access

Sources

Source	URL	Tier
Security News	https://thehackernews.com/2026/06/new-attacks-trick-openclaw-ai-age...	T3
OpenClaw 2026.4.21 released with OpenAI Image 2 and security fixes	https://www.facebook.com/groups/artificialintelligenceforbusines/po...	T3
Latest Features & Release Notes - OpenClaw Updates	https://openclaw.com.au/updates	T3
Releases · openclaw/openclaw - GitHub	https://github.com/openclaw/openclaw/releases	T3
OpenClaw 2026.5.26 update with faster replies and meeting notes	https://www.facebook.com/groups/9010lifestyle/posts/3892312197570810/	T3

Generated 2026-06-16 08:12 UTC by TJS Security Command Center