Se ha dado a conocer una vulnerabilidad de seguridad de alta gravedad en OpenClaw (anteriormente conocido como Clawdbot y Moltbot) que podría permitir la ejecución remota de código (RCE) mediante un enlace malicioso especialmente diseñado.

El fallo, identificado como CVE-2026-25253 (puntuación CVSS: 8.8), fue corregido en la versión 2026.1.29 publicada el 30 de enero de 2026. Se trata de una vulnerabilidad de exfiltración de tokens que puede derivar en la toma de control total del gateway.

"La interfaz de Control confía en el parámetro gatewayUrl proveniente de la cadena de consulta sin validarlo y se conecta automáticamente al cargarse, enviando el token del gateway almacenado en la carga de conexión WebSocket", explicó Peter Steinberger, creador y mantenedor de OpenClaw, en un aviso de seguridad.

"Al hacer clic en un enlace manipulado o visitar un sitio malicioso, el token puede enviarse a un servidor controlado por un atacante. Posteriormente, el atacante puede conectarse al gateway local de la víctima, modificar la configuración (sandbox, políticas de herramientas) e invocar acciones privilegiadas, logrando una RCE con un solo clic".

OpenClaw es un asistente personal de inteligencia artificial (IA) autónomo y de código abierto que se ejecuta localmente en los dispositivos de los usuarios e integra múltiples plataformas de mensajería. Aunque fue lanzado inicialmente en noviembre de 2025, el proyecto ha experimentado un crecimiento acelerado en las últimas semanas, y su repositorio en GitHub ya supera las 149.000 estrellas al momento de redactar este informe.

"OpenClaw es una plataforma de agentes abiertos que se ejecuta en tu propia máquina y funciona desde las aplicaciones de chat que ya utilizas", afirmó Steinberger. *"A diferencia de los asistentes SaaS, donde tus datos residen en servidores de terceros, OpenClaw se ejecuta donde tú decidas: portátil, homelab o VPS. Tu infraestructura. Tus claves. Tus datos."*

Mav Levin, investigador de seguridad fundador de depthfirst y a quien se atribuye el descubrimiento de la falla, señaló que esta puede explotarse para crear una cadena de ataque RCE de un solo clic que se ejecuta en cuestión de milisegundos tras la visita de la

víctima a una única página web maliciosa.

El problema radica en que basta con hacer clic en el enlace hacia dicha página para activar un ataque de secuestro de WebSocket entre sitios, ya que el servidor de OpenClaw no valida el encabezado de origen del WebSocket. Esto permite que el servidor acepte solicitudes desde cualquier sitio web, eludiendo de facto las restricciones de red asociadas a localhost.

Una página web maliciosa puede aprovechar esta debilidad para ejecutar JavaScript del lado del cliente en el navegador de la víctima, recuperar un token de autenticación, establecer una conexión WebSocket con el servidor y utilizar el token robado para evadir la autenticación e iniciar sesión en la instancia de OpenClaw del usuario afectado.

La situación se agrava aún más porque, al aprovechar los permisos privilegiados operator.admin y operator.approvals asociados al token, el atacante puede usar la API para desactivar la confirmación del usuario estableciendo “exec.approvals.set” en “off” y escapar del contenedor que ejecuta las herramientas de shell configurando “tools.exec.host” como “gateway”.

“Esto obliga al agente a ejecutar comandos directamente en la máquina anfitriona, y no dentro de un contenedor Docker”, explicó Levin. *“Finalmente, para lograr la ejecución arbitraria de comandos, el JavaScript del atacante ejecuta una solicitud node.invoke.”*

Consultado sobre si el uso de la API de OpenClaw para gestionar las funciones de seguridad representa una limitación arquitectónica, Levin comentó que *“el problema es que estas defensas (sandbox y barreras de seguridad) fueron diseñadas para contener acciones maliciosas de un LLM, por ejemplo como resultado de una inyección de prompts. Los usuarios podrían pensar que estas protecciones los resguardan de esta vulnerabilidad o limitan su impacto, pero no es así.”*

Steinberger indicó en el aviso que *“la vulnerabilidad puede explotarse incluso en instancias configuradas para escuchar únicamente en loopback, ya que el navegador de la víctima es quien inicia la conexión saliente.”*

“Afecta a cualquier implementación de Moltbot en la que un usuario se haya autenticado en la interfaz de Control. El atacante obtiene acceso a nivel operador a la API del gateway, lo que permite cambios arbitrarios en la configuración y la ejecución de código en el host del gateway. El ataque funciona incluso cuando el gateway está vinculado a loopback, porque el navegador de la víctima actúa como intermediario.”