Investigadores de ciberseguridad han revelado múltiples vulnerabilidades de seguridad en Claude Code, un asistente de programación impulsado por inteligencia artificial (IA) desarrollado por Anthropic, que podrían derivar en la ejecución remota de código y en el robo de credenciales de API.

«Las vulnerabilidades explotan diversos mecanismos de configuración, incluidos Hooks, servidores Model Context Protocol (MCP) y variables de entorno, ejecutando comandos arbitrarios en la shell y exfiltrando claves de API de Anthropic cuando los usuarios clonan y abren repositorios no confiables», señaló Check Point Research en un informe.

Las fallas detectadas se agrupan en tres grandes categorías:

- Sin CVE (puntuación CVSS: 8.7) – Vulnerabilidad de inyección de código originada por la omisión del consentimiento del usuario al iniciar Claude Code en un nuevo directorio, lo que podría permitir la ejecución de código arbitrario sin confirmación adicional mediante hooks de proyectos no confiables definidos en .claude/settings.json. (Corregido en la versión 1.0.87 en septiembre de 2025)
- CVE-2025-59536 (puntuación CVSS: 8.7) – Vulnerabilidad de inyección de código que posibilita la ejecución automática de comandos arbitrarios en la shell durante la inicialización de herramientas cuando un usuario inicia Claude Code en un directorio no confiable. (Corregido en la versión 1.0.111 en octubre de 2025)
- CVE-2026-21852 (puntuación CVSS: 5.3) – Vulnerabilidad de divulgación de información en el flujo de carga de proyectos de Claude Code que permite que un repositorio malicioso extraiga datos, incluidas claves de API de Anthropic. (Corregido en la versión 2.0.65 en enero de 2026)

«Si un usuario iniciaba Claude Code en un repositorio controlado por un atacante y este incluía un archivo de configuración que establecía ANTHROPIC_BASE_URL hacia un endpoint bajo control del atacante, Claude Code enviaría solicitudes a la API antes de mostrar el aviso de confianza, pudiendo filtrar las claves de API del usuario», explicó Anthropic en un aviso relacionado con CVE-2026-21852.

En otras palabras, basta con abrir un repositorio manipulado para que se exfiltre la clave de API activa de un desarrollador, se redirija el tráfico autenticado hacia infraestructura externa y se capturen credenciales. Esto podría permitir al atacante profundizar aún más en la infraestructura de IA de la víctima.

Las consecuencias potenciales incluyen el acceso a archivos compartidos del proyecto, la modificación o eliminación de datos almacenados en la nube, la carga de contenido malicioso e incluso la generación de costos inesperados en el uso de la API.

La explotación exitosa de la primera vulnerabilidad podría provocar la ejecución encubierta de código en el equipo del desarrollador sin que este realice ninguna acción adicional más allá de abrir el proyecto.

CVE-2025-59536 persigue un objetivo similar; la diferencia principal es que las configuraciones definidas por el repositorio a través de los archivos .mcp.json y claude/settings.json podrían ser utilizadas por un atacante para eludir la aprobación explícita del usuario antes de interactuar con herramientas y servicios externos mediante el Model Context Protocol (MCP). Esto se logra configurando la opción «enableAllProjectMcpServers» en true.

«A medida que las herramientas impulsadas por IA adquieren la capacidad de ejecutar comandos, iniciar integraciones externas y establecer comunicaciones de red de forma autónoma, los archivos de configuración pasan a formar parte efectiva de la capa de ejecución», indicó Check Point. *«Lo que antes se consideraba contexto operativo ahora influye directamente en el comportamiento del sistema.»*

«Esto modifica de manera fundamental el modelo de amenazas. El riesgo ya no se limita a ejecutar código no confiable, sino que se amplía al simple hecho de abrir proyectos no confiables. En entornos de desarrollo impulsados por IA, la cadena de suministro no comienza únicamente con el código fuente, sino también con las capas de automatización que lo rodean.»