OpenClaw (antes conocido como Moltbot y Clawdbot) anunció que se ha asociado con VirusTotal, propiedad de Google, para analizar las skills que se suben a ClawHub, su mercado de habilidades, como parte de un esfuerzo más amplio para reforzar la seguridad del ecosistema de agentes autónomos.

"Todas las skills publicadas en ClawHub ahora se analizan utilizando la inteligencia de amenazas de VirusTotal, incluida su nueva capacidad Code Insight", afirmaron el fundador de OpenClaw, Peter Steinberger, junto con Jamieson O'Reilly y Bernardo Quintero. *"Esto añade una capa adicional de protección para la comunidad de OpenClaw."*

El procedimiento consiste básicamente en generar un hash SHA-256 único para cada skill y compararlo con la base de datos de VirusTotal en busca de coincidencias. Si no se encuentra ningún resultado, el paquete de la skill se sube a la herramienta de análisis de malware para un examen más profundo mediante VirusTotal Code Insight.

Las skills que reciben un veredicto *"benigno"* por parte de Code Insight se aprueban automáticamente en ClawHub, mientras que aquellas catalogadas como sospechosas se marcan con una advertencia. Cualquier skill considerada maliciosa se bloquea y no puede descargarse. OpenClaw también indicó que todas las skills activas se vuelven a analizar diariamente para detectar casos en los que una skill previamente limpia pase a ser maliciosa.

No obstante, los responsables de OpenClaw advirtieron que el escaneo con VirusTotal *"no es una bala de plata"* y que existe la posibilidad de que algunas skills maliciosas, que empleen cargas de inyección de prompts cuidadosamente ocultas, logren evadir los controles.

Además de la colaboración con VirusTotal, la plataforma tiene previsto publicar un modelo de amenazas integral, una hoja de ruta de seguridad pública, un proceso formal de reporte de vulnerabilidades y detalles sobre la auditoría de seguridad de todo su código base.

Este avance se produce tras informes que detectaron cientos de skills maliciosas en ClawHub, lo que llevó a OpenClaw a añadir una opción de reporte que permite a los usuarios autenticados señalar habilidades sospechosas. Diversos análisis han revelado que estas skills

se hacen pasar por herramientas legítimas, pero en realidad incorporan funcionalidades maliciosas destinadas a exfiltrar datos, insertar puertas traseras para acceso remoto o instalar malware de tipo stealer.

"Los agentes de IA con acceso al sistema pueden convertirse en canales encubiertos de fuga de datos que eluden las herramientas tradicionales de prevención de pérdida de datos, los proxies y la monitorización de endpoints", señaló Cisco la semana pasada. *"Además, los modelos también pueden actuar como orquestadores de ejecución, donde el propio prompt se convierte en la instrucción y resulta difícil de detectar con herramientas de seguridad convencionales."*

La reciente popularidad viral de OpenClaw, el asistente de inteligencia artificial de código abierto basado en agentes, y de Moltbook, una red social adyacente donde agentes autónomos creados sobre OpenClaw interactúan entre sí en una plataforma al estilo Reddit, ha despertado preocupaciones en materia de seguridad.

Si bien OpenClaw funciona como un motor de automatización para activar flujos de trabajo, interactuar con servicios en línea y operar en distintos dispositivos, el amplio nivel de acceso concedido a las skills, sumado a que pueden procesar datos procedentes de fuentes no confiables, abre la puerta a riesgos como malware e inyección de prompts.

En otras palabras, estas integraciones, aunque prácticas, amplían de forma considerable la superficie de ataque y el conjunto de entradas no confiables que consume el agente, convirtiéndolo en un *"caballo de Troya agentico"* para la exfiltración de datos y otras acciones maliciosas. Backslash Security ha descrito a OpenClaw como una *"IA con manos"*.

"A diferencia del software tradicional, que hace exactamente lo que el código le indica, los agentes de IA interpretan lenguaje natural y toman decisiones sobre acciones", señaló OpenClaw. *"Difuminan la frontera entre la intención del usuario y la ejecución de la máquina. Pueden ser manipulados a través del propio lenguaje."*

OpenClaw también reconoció que el poder de las skills —utilizadas para ampliar las

capacidades de un agente de IA, desde controlar dispositivos del hogar inteligente hasta gestionar finanzas— puede ser explotado por actores maliciosos, que pueden aprovechar el acceso del agente a herramientas y datos para extraer información sensible, ejecutar comandos no autorizados, enviar mensajes en nombre de la víctima e incluso descargar y ejecutar cargas adicionales sin su conocimiento ni consentimiento.

Además, dado que OpenClaw se está desplegando cada vez más en endpoints corporativos sin la aprobación formal de TI o de seguridad, los privilegios elevados de estos agentes pueden facilitar aún más el acceso a la shell, el movimiento de datos y la conectividad de red fuera de los controles de seguridad estándar, creando una nueva categoría de riesgo de *Shadow AI* para las empresas.

"OpenClaw y herramientas similares aparecerán en tu organización te guste o no", afirmó Tomer Yahalom, investigador de Astrix Security. *"Los empleados las instalarán porque realmente son útiles. La única incógnita es si tú lo sabrás."*

Algunos de los problemas de seguridad más evidentes que han salido a la luz recientemente se enumeran a continuación:

- Un problema ya corregido en versiones anteriores que podía provocar que el tráfico proxy se clasificara erróneamente como local, omitiendo la autenticación en algunas instancias expuestas a Internet.
- *"OpenClaw almacena credenciales en texto plano, utiliza patrones de codificación inseguros como eval directo con entrada del usuario y no cuenta con una política de privacidad ni una rendición de cuentas clara"*, indicaron Moshe Siman Tov Bustan y Nir Zadok, de OX Security. *"Los métodos comunes de desinstalación dejan datos sensibles atrás, y revocar completamente el acceso es mucho más difícil de lo que la mayoría de los usuarios cree."*
- Un ataque de cero clic que abusa de las integraciones de OpenClaw para implantar una puerta trasera en el endpoint de la víctima cuando el agente de IA procesa un documento aparentemente inofensivo, ejecutando una carga de inyección de prompt indirecta que permite responder a mensajes desde un bot de Telegram controlado por

el atacante.

- Una inyección de prompt indirecta incrustada en una página web que, al ser analizada como parte de un prompt legítimo que pide al modelo de lenguaje resumir el contenido, provoca que OpenClaw añada instrucciones controladas por el atacante al archivo `~/.openclaw/workspace/HEARTBEAT.md` y espere silenciosamente nuevas órdenes desde un servidor externo.
- Un análisis de seguridad de 3,984 skills en el mercado ClawHub detectó que 283 skills, aproximadamente el 7.1 % del registro total, contienen fallas críticas que exponen credenciales sensibles en texto plano a través del contexto y los registros de salida del LLM.
- Un informe de Bitdefender reveló que las skills maliciosas suelen clonarse y republicarse a gran escala con pequeñas variaciones en el nombre, y que las cargas se alojan mediante servicios de paste como glot.io y repositorios públicos de GitHub.
- Una vulnerabilidad de ejecución remota de código, ya parcheada, que podía permitir a un atacante engañar a un usuario para visitar una página web maliciosa capaz de hacer que la interfaz Gateway Control filtrara el token de autenticación de OpenClaw a través de un canal WebSocket y luego usarlo para ejecutar comandos arbitrarios en el host.
- El gateway de OpenClaw se enlaza por defecto a `0.0.0.0:18789`, exponiendo la API completa a cualquier interfaz de red. Según datos de Censys, existen más de 30,000 instancias expuestas accesibles por Internet al 8 de febrero de 2026, aunque la mayoría requiere un token para poder interactuar con ellas.
- En un escenario de ataque hipotético, una carga de inyección de prompt incrustada en un mensaje de WhatsApp especialmente diseñado puede utilizarse para exfiltrar los archivos `.env` y `creds.json`, que almacenan credenciales, claves API y tokens de sesión de plataformas de mensajería conectadas desde una instancia expuesta de OpenClaw.
- Una base de datos de Supabase mal configurada perteneciente a Moltbook quedó expuesta en JavaScript del lado del cliente, dejando accesibles las claves API secretas de todos los agentes registrados en el sitio y permitiendo acceso completo de lectura y escritura a los datos de la plataforma. Según Wiz, la filtración incluyó 1.5 millones de tokens de autenticación API, 35,000 direcciones de correo electrónico y mensajes

privados entre agentes.
- Se ha observado a actores de amenazas explotando la mecánica de la plataforma Moltbook para amplificar su alcance y dirigir a otros agentes hacia hilos maliciosos que contienen inyecciones de prompts destinadas a manipular su comportamiento y extraer datos sensibles o robar criptomonedas.
- *"Moltbook puede haber creado inadvertidamente un laboratorio en el que los agentes, que pueden ser objetivos de alto valor, procesan e interactúan constantemente con datos no confiables, y donde las barreras de protección no están integradas en la plataforma, todo por diseño"*, señaló Zenity Labs.
- *"El primer problema, y quizá el más grave, es que OpenClaw depende del modelo de lenguaje configurado para muchas decisiones críticas de seguridad"*, indicaron los investigadores de HiddenLayer Conor McCauley, Kasimir Schulz, Ryan Tracey y Jason Martin. *"A menos que el usuario habilite proactivamente la función de aislamiento de herramientas basada en Docker, el acceso completo al sistema sigue siendo el valor por defecto."*

Entre otros problemas de arquitectura y diseño identificados por la empresa de seguridad en IA se encuentran la incapacidad de OpenClaw para filtrar contenido no confiable que contiene secuencias de control, protecciones ineficaces contra inyecciones de prompts indirectas, memorias y prompts del sistema modificables que persisten entre sesiones, almacenamiento en texto plano de claves API y tokens de sesión, y la ausencia de una aprobación explícita del usuario antes de ejecutar llamadas a herramientas.

En un informe publicado la semana pasada, Persmiso Security sostuvo que la seguridad del ecosistema OpenClaw es aún más crítica que la de las tiendas de aplicaciones y los mercados de extensiones de navegador, debido al amplio acceso que estos agentes tienen a los datos del usuario.

"Los agentes de IA obtienen credenciales de toda tu vida digital", advirtió el investigador de seguridad Ian Ahl. *"Y, a diferencia de las extensiones de navegador que se ejecutan en un entorno aislado con cierto nivel de separación, estos agentes operan con todos los privilegios que les otorgas."*

“El mercado de skills agrava este problema. Cuando instalas una extensión de navegador maliciosa, comprometes un solo sistema. Cuando instalas una skill maliciosa de un agente, potencialmente comprometes todos los sistemas para los que ese agente tiene credenciales.”

La extensa lista de problemas de seguridad asociados con OpenClaw llevó al Ministerio de Industria y Tecnología de la Información de China a emitir una alerta sobre instancias mal configuradas, instando a los usuarios a implementar medidas de protección frente a ciberataques y filtraciones de datos, según informó Reuters.

“Cuando las plataformas de agentes se vuelven virales más rápido de lo que maduran las prácticas de seguridad, la mala configuración se convierte en la principal superficie de ataque”, afirmó Ensar Seker, CISO de SOCRadar. *“El riesgo no es el agente en sí, sino exponer herramientas autónomas a redes públicas sin identidades reforzadas, controles de acceso y límites claros de ejecución.”*

“Lo destacable aquí es que el regulador chino está señalando explícitamente el riesgo de configuración en lugar de prohibir la tecnología. Esto coincide con lo que los defensores ya saben: los frameworks de agentes amplifican tanto la productividad como el impacto de un incidente. Un único endpoint expuesto o un plugin excesivamente permisivo puede convertir a un agente de IA en una capa de automatización involuntaria para los atacantes.”