Dos vulnerabilidades de seguridad identificadas en Paperclip podrían permitir que un atacante ejecute comandos tanto en un servidor de red como en el equipo de un desarrollador. Paperclip es un plano de control (*control plane*) de código abierto diseñado para administrar equipos de agentes de inteligencia artificial (IA), y ambos vectores de ataque requieren importar un agente malicioso y ponerlo en funcionamiento.

Una tercera vulnerabilidad podría exponer información confidencial y detalles del plano de control a través de rutas de la interfaz de programación de aplicaciones (API) que no implementaban los mecanismos de control de acceso previstos.

La vulnerabilidad más crítica, identificada como CVE-2026-41679 (CVSS: 10.0), puede explotarse sin necesidad de disponer previamente de una cuenta ni de interacción por parte de la víctima, siempre que la implementación sea accesible desde la red, opere en modo autenticado y conserve la configuración predeterminada de registro.

La segunda vulnerabilidad, registrada como GHSA-x8hx-rhr2-9rf7 (CVSS: 9.6), requiere que el usuario visite una página controlada por el atacante mientras Paperclip se encuentra ejecutándose en su configuración predeterminada local_trusted.

El código fuente correspondiente a Paperclip v2026.416.0 incorpora las correcciones relacionadas con la autorización durante la importación y la validación del nombre de host descritas posteriormente. No obstante, el aviso referente al ataque de DNS rebinding no especifica explícitamente una versión corregida. Posteriormente, Rapid7 publicó un módulo para Metasploit que automatiza la explotación de CVE-2026-41679, mientras que la clasificación Stakeholder-Specific Vulnerability Categorization (SSVC) incorporada por NVD considera que la explotación se encuentra demostrada mediante una prueba de concepto (*proof of concept*).

Hasta el 5 de agosto de 2026, ninguna de las fuentes oficiales revisadas por *The Hacker News* había confirmado ataques activos que explotaran estas vulnerabilidades. Por ello, se recomienda actualizar Paperclip a la versión v2026.416.0 o superior y revisar cuidadosamente la configuración del registro de usuarios y la exposición de las

implementaciones.

El análisis realizado por Oasis Security, respaldado por un informe técnico de 17 páginas, concluye que todas las vulnerabilidades comparten una característica común: la configuración de los agentes puede transformarse en comportamiento ejecutable. El adaptador de procesos integrado en Paperclip está diseñado para ejecutar el comando configurado como un proceso hijo del servidor.

Esta funcionalidad constituye una característica legítima del producto. Sin embargo, las vulnerabilidades modificaban quién podía acceder a dicha capacidad y qué configuraciones eran consideradas confiables por el servidor.

«La configuración de los agentes debe tratarse como una entrada ejecutable«, señaló Oasis.

Como consecuencia, usuarios no autorizados o solicitudes originadas desde un navegador podían introducir y activar configuraciones que finalmente eran ejecutadas por el componente encargado de iniciar procesos.

La cadena de ataque dirigida al servidor afecta implementaciones autenticadas accesibles desde Internet que mantienen la configuración vulnerable de registro. En contraste, el vector dirigido al entorno local únicamente es posible cuando un usuario accede a una página controlada por un atacante mientras Paperclip funciona en modo local_trusted. En ambos casos, el resultado final consiste en que una configuración de agente controlada por el atacante alcanza el adaptador responsable de ejecutar comandos en el sistema anfitrión.

Una clave de tablero aprobada por su propio propietario

El ataque contra una instancia expuesta en Internet comienza aprovechando el mecanismo predeterminado de registro abierto de Paperclip. Un atacante puede crear una cuenta sin necesidad de invitación ni de verificar una dirección de correo electrónico, iniciar sesión y acceder al proceso de autorización de la interfaz de línea de comandos (CLI). Posteriormente, el mismo usuario recién registrado puede generar una solicitud de autorización pendiente

para la CLI y aprobarla por sí mismo, obteniendo así una credencial permanente para la API del tablero sin requerir la aprobación de un administrador.

En condiciones normales, dicha credencial no debería otorgar privilegios suficientes para crear una nueva empresa (*company*). Paperclip exigía permisos de administrador de la instancia cuando la empresa era creada directamente, pero la ruta de importación equivalente aceptaba únicamente privilegios de nivel de tablero.

Gracias a esta inconsistencia, un atacante podía importar un archivo .paperclip.yaml que definiera una nueva empresa, un agente basado en el adaptador de procesos y el comando que dicho agente debía ejecutar.

El propio proceso de importación añadía automáticamente al atacante como miembro de la nueva empresa, permitiendo superar la comprobación de activación (*wakeup*) cuando el agente era iniciado. Como resultado, Paperclip ejecutaba el comando utilizando los privilegios del proceso del servidor sobre el sistema operativo.

El impacto práctico dependía de los privilegios asignados a la cuenta de servicio y de la configuración del sistema anfitrión. Según Oasis, esto podía implicar el acceso a datos de la aplicación, repositorios de código fuente, credenciales almacenadas localmente, secretos disponibles para los agentes y servicios internos accesibles desde el servidor.

La vulnerabilidad CVE-2026-41679 fue corregida en la versión v2026.416.0, exigiendo permisos de administrador para importar nuevas empresas y permisos específicos de empresa cuando la importación se dirige a organizaciones ya existentes. Esta misma validación protege ahora tanto la vista previa de la importación como su ejecución definitiva.

Aunque el registro abierto continúa disponible, los usuarios recién registrados ya no pueden utilizar la función de importación de nuevas empresas para obtener privilegios equivalentes a los de un administrador de instancia.

En junio de 2026, Rapid7 publicó un módulo para Metasploit que automatiza la cadena

completa de explotación de CVE-2026-41679, compuesta por seis solicitudes. Asimismo, la evaluación realizada por CISA-ADP, reflejada en NVD, clasifica la vulnerabilidad como automatizable, de impacto técnico total y con prueba de concepto disponible.

No obstante, al 5 de agosto de 2026, esta vulnerabilidad todavía no figuraba en el catálogo Known Exploited Vulnerabilities (KEV) de CISA. Su ausencia en dicho listado no implica necesariamente que no haya sido explotada en entornos reales.

El navegador cruza hacia localhost

La segunda vulnerabilidad crítica afecta un modelo de implementación diferente. En la configuración predeterminada local_trusted, Paperclip escucha únicamente en la interfaz de *loopback* y, tradicionalmente, consideraba que cualquier solicitud recibida desde esa interfaz pertenecía implícitamente a un administrador de la instancia. Esta decisión simplificaba el desarrollo local, pero asumía erróneamente que la ubicación de red equivalía a una identidad confiable.

Oasis demostró un ataque de DNS rebinding mediante un nombre de dominio controlado por el atacante que resolvía tanto hacia su propio servidor como hacia la dirección 127.0.0.1. Inicialmente, el navegador descargaba código JavaScript desde el servidor malicioso. Posteriormente, cuando este dejaba de responder, las solicitudes dirigidas al mismo dominio eran redirigidas al servicio local de Paperclip.

El navegador continuaba considerando dichas solicitudes como pertenecientes al mismo origen (*same-origin*), mientras que Paperclip aceptaba el nombre del host proporcionado en el encabezado Host. De esta manera, la página maliciosa podía invocar la API de importación, instalar una empresa que contenía un agente basado en procesos y activar posteriormente su punto de ejecución (*wakeup*).

Dado que el modo local otorgaba privilegios administrativos a las solicitudes resultantes del ataque de *rebinding*, el servidor ejecutaba el comando del atacante con los permisos del desarrollador. Para lograrlo no era necesario disponer de tokens de Paperclip, cookies de

sesión ni credenciales robadas. La prueba de concepto documentada fue validada en macOS utilizando Firefox, aunque no existe evidencia pública que confirme el mismo comportamiento en todos los navegadores y sistemas operativos.

La corrección implementada consistió en incorporar una validación estricta del nombre de host. En la versión v2026.416.0, Paperclip habilita un mecanismo de protección para implementaciones privadas que operan tanto en modo local_trusted como en modo autenticado. Esta validación se ejecuta antes del componente responsable de asignar la identidad de la solicitud, rechazando cualquier petición cuyo nombre de host no haya sido previamente autorizado.

Rutas API sin mecanismos de protección

La tercera vulnerabilidad, identificada como GHSA-xfqj-r5qw-8g4j (CVSS: 8.3), afecta diversas rutas de la API disponibles en modo autenticado que no verificaban correctamente si las solicitudes provenían de usuarios autenticados o pertenecientes a la empresa correspondiente.

Un usuario que conociera un identificador válido asociado a una ejecución (*heartbeat-run identifier*) podía recuperar información vinculada a incidencias sin demostrar que poseía permisos sobre la empresa correspondiente. Aunque esta vulnerabilidad no permitía una enumeración masiva de datos, sí ofrecía una vía de divulgación dirigida para quienes lograran obtener identificadores válidos.

Otras rutas expuestas devolvían documentación interna destinada a los agentes de Paperclip, incluyendo rutas de la API y convenciones de autenticación, además de información sobre el estado de salud del sistema, como el modo de implementación, la versión instalada, el estado de autenticación, la configuración inicial (*bootstrap*), el nivel de exposición y las funcionalidades habilitadas. Asimismo, la ruta no autenticada utilizada para los desafíos de la CLI formaba parte de la cadena de generación de credenciales aprovechada durante la explotación de CVE-2026-41679.

El problema fundamental residía en el diseño del sistema, que permitía que solicitudes no autenticadas atravesaran el *middleware* con una identidad equivalente a «sin actor», delegando en cada ruta la responsabilidad de aplicar sus propias comprobaciones de seguridad. Como respuesta, Paperclip incorporó autenticación obligatoria para las rutas generales de habilidades (*skills*), verificaciones de acceso a nivel de empresa para la recuperación de incidencias, controles específicos para los procesos de incorporación mediante invitaciones y una respuesta reducida de los servicios de estado cuando las solicitudes no se encuentran autenticadas.

En conjunto, las tres vulnerabilidades reflejan un patrón común: Paperclip depositaba confianza en una credencial, una ruta o una ubicación de red sin aplicar las validaciones requeridas por la operación posterior. Entre los fallos específicos destacan la autoaprobación de credenciales, la ausencia de controles de acceso en determinadas rutas y la suposición de que toda conexión proveniente de localhost era inherentemente confiable.

Inconsistencias en el registro de versiones

Paperclip utiliza dos esquemas de numeración distintos para identificar el mismo código fuente. Mientras que la publicación de seguridad en GitHub hace referencia a la versión v2026.416.0, los manifiestos correspondientes al servidor y a la interfaz CLI incluidos en esa misma etiqueta indican la versión 0.3.1. Esta doble nomenclatura explica que los distintos avisos de seguridad mencionen indistintamente ambas versiones.

El aviso correspondiente al ataque de DNS rebinding continúa sin especificar una versión corregida. Sin embargo, una revisión realizada por *The Hacker News* del código etiquetado como v2026.416.0 confirmó que la validación del nombre de host ya se encuentra habilitada para implementaciones privadas en modo local_trusted. Asimismo, esta versión exige privilegios de administrador para importar nuevas empresas y bloquea el uso de los adaptadores de procesos y HTTP dentro de la ruta restringida de importación segura de agentes.

A pesar de estas correcciones, persisten inconsistencias en la documentación. Mientras el

código de v2026.416.0 incorpora las medidas de mitigación necesarias, el aviso sobre DNS rebinding continúa sin indicar una versión corregida y NVD mantiene en su historial información correspondiente a versiones afectadas anteriores. En consecuencia, la recomendación más segura para los administradores consiste en actualizar directamente a v2026.416.0 o una versión posterior, en lugar de basarse exclusivamente en los metadatos históricos.

Finalmente, las notas oficiales de lanzamiento de Paperclip recomiendan que todas las implementaciones sean actualizadas, recomendación que también respalda Oasis Security al sugerir el uso de la versión v2026.416.0 o superior.