Oscar Obando
CHAVES

Un clic y tu correo se fue

Un enlace bastaba para que Copilot vaciara tu Gmail, tu Drive y tu calendario sin pedirte confirmación. Y la instrucción que dejaba escrita en su memoria sobrevivía a que cambiaras la contraseña.

Oscar Obando Chaves

oscarobando.ai

Agosto de 2026

LA FALLA

Lo que ocurría al hacer clic en un enlace

El 18 de agosto el equipo de investigación de la firma Varonis publicó el detalle de una cadena de tres vulnerabilidades en la versión personal de Microsoft Copilot que permitía robar información de las cuentas conectadas de una víctima con un solo clic, y la mecánica merece conocerse porque describe un tipo de ataque que va a repetirse en cualquier asistente con acceso a tus cuentas.

El conjunto recibió el nombre de **CoSnitch** y quedó registrado como **CVE-2026-24301** con una calificación de severidad de **8,8 sobre 10** en la escala estándar de la industria, y su pieza central era un parámetro de dirección web sin documentar que hacía que el asistente ejecutara de forma automática la instrucción incrustada en el enlace, sin ninguna confirmación visible para quien acababa de hacer clic.

Lo que venía después ocurría dentro de tu propia sesión ya autenticada, porque la instrucción del atacante operaba con los permisos que tú mismo habías concedido, y desde ahí podía recuperar información de los servicios conectados como **Gmail**, **Google Drive** y el calendario, además de la memoria del asistente y el historial completo de tus conversaciones con él.

El punto que vuelve difícil de detectar a este ataque lo describieron los propios investigadores al señalar que los sistemas de vigilancia de seguridad veían al asistente haciendo exactamente lo que hace siempre, que es consultar una dirección web, de modo que la extracción de datos viajaba disfrazada de una petición rutinaria para resumir una página cualquiera.

Fuentes. Varonis Threat Labs, publicación del 18 de agosto de 2026. Microsoft Security Update Guide, CVE-2026-24301, calificación 8,8 bajo CVSS 3.1. Cobertura de Dark Reading, The Hacker News y TechRadar (18 al 20 de agosto de 2026).

La parte que sobrevive al cambio de contraseña

El tercer eslabón de la cadena es el que conviene entender con más cuidado, porque introduce una forma de persistencia que rompe casi todas las reacciones que una persona tomaría al sospechar que algo pasó con su cuenta.

El mecanismo consistía en alojar una página aparentemente inofensiva con instrucciones escondidas dentro de comentarios del código, en la metadata o en elementos visualmente ocultos, de manera que cuando la víctima le pedía al asistente un resumen de esa página, el sistema procesaba tanto el texto visible como las instrucciones del atacante y las escribía en su **memoria permanente**.

La advertencia de los investigadores sobre esa persistencia es literal y conviene citarla completa, porque señalaron que la inyección sobrevive a los cambios de contraseña, a la revocación de sesiones y a la reinscripción del dispositivo, o sea a las tres medidas que cualquier persona toma cuando sospecha que alguien entró a su cuenta.

La consecuencia práctica es que las instrucciones inyectadas quedan activas en sesiones futuras hasta que alguien las elimine de forma manual, y la propia divulgación de Varonis registra un punto pendiente que conviene tener presente, ya que no aclara si la corrección de Microsoft eliminó de forma retroactiva las entradas de memoria creadas antes del parche.

Fuentes. Varonis Threat Labs (18 de agosto de 2026), sobre la persistencia de la inyección en memoria. The Hacker News (agosto de 2026), sobre el punto no aclarado respecto de la remediación retroactiva. TechRadar (agosto de 2026).

OCHO MESES

Cuánto tiempo estuvo abierta la puerta

La línea de tiempo del caso plantea una discusión que excede a este fallo concreto, porque el intervalo entre el aviso y la corrección determina cuánto tiempo estuvieron expuestos los usuarios sin saberlo ni poder hacer nada al respecto.

Varonis reportó el problema a Microsoft en **diciembre de 2025** mediante un proceso de divulgación coordinada, y el parche llegó el **18 de agosto de 2026**, de modo que transcurrieron alrededor de **ocho meses** entre el aviso y la corrección, un período durante el cual la vulnerabilidad permaneció activa en el producto.

La posición de Microsoft matiza el alcance del problema y merece registrarse con precisión, porque la empresa señaló que el fallo afectaba únicamente a la versión personal del asistente, que los clientes empresariales quedaron fuera del alcance, y que no se requiere ninguna acción del usuario porque la corrección ya está desplegada del lado del servidor.

Varonis por su parte declaró que no encontró evidencia de que el fallo haya sido explotado en la práctica antes del parche, y añadió un dato que ordena la lectura del caso, porque este es el tercer fallo de este mismo asistente que la firma reporta durante el año, después de otros dos anteriores, lo que sugiere una superficie de ataque que todavía está siendo cartografiada.

Fuentes. Varonis Threat Labs (18 de agosto de 2026) sobre la línea de tiempo de la divulgación coordinada. Declaraciones de Microsoft a Dark Reading (agosto de 2026). Redmondmag (19 de agosto de 2026).

POR QUÉ SE REPITE

La confusión que ningún asistente ha resuelto

Este fallo pertenece a una familia de ataques que aparece una y otra vez en productos distintos de empresas distintas, y el motivo es que todos comparten una limitación de fondo que sigue sin solución técnica satisfactoria.

La raíz se conoce como **inyección indirecta de instrucciones** y consiste en que estos sistemas siguen sin distinguir de forma confiable entre el contenido que deben analizar y las órdenes que deben obedecer, así que un texto escondido dentro de una página, de un documento o de un correo puede ser leído como una instrucción legítima del usuario.

El patrón se repitió durante este año en varios productos, porque otros investigadores documentaron por separado escrituras y borrados de memoria mediante la misma técnica en la versión empresarial del mismo asistente, y la propia firma que reportó este caso ya había publicado antes otros dos fallos distintos del mismo producto durante el año.

Lo que agrava el asunto en el terreno de los agentes es la acumulación de accesos, ya que cada conector activo amplía lo que una instrucción inyectada puede alcanzar, de modo que un asistente conectado al correo, al almacenamiento y al calendario concentra detrás de una sola conversación el acceso a información que antes vivía repartida en tres sistemas separados con controles independientes.

Fuentes. Investigaciones de Håkon Måløy y Johann Rehberger sobre inyección indirecta y modificación de memoria en asistentes de Microsoft (junio de 2026), asociadas a CVE-2026-24299. Varonis Threat Labs, reportes previos del mismo producto durante 2026.

LA CADENA COMPLETA

Los tres eslabones y qué hacía cada uno

Separada en sus tres piezas, la cadena muestra que ninguna de ellas por sí sola habría bastado y que el daño surgió de encadenarlas.

Eslabón	Qué hacía	Por qué funcionaba
Parámetro sin documentar	**Ejecución automática**	El asistente corría la instrucción del enlace sin pedir confirmación al usuario.
Abuso de permisos concedidos	**Extracción de datos**	Operaba dentro de la sesión ya autenticada, con los accesos que el usuario había otorgado.
Inyección indirecta al resumir	**Envenenamiento de memoria**	Instrucciones ocultas en una página quedaban escritas en la memoria permanente.

El segundo eslabón contiene la lección estructural del caso, porque el ataque jamás vulneró el sistema de autorización ni la cuenta de correo, y se limitó a usar los permisos que la víctima había concedido de forma legítima, de modo que la falla conceptual estuvo en presumir que toda invocación de un servicio conectado la inicia una persona de forma deliberada.

Fuentes. Varonis Threat Labs (18 de agosto de 2026). Redmondmag y gbhackers (agosto de 2026), sobre la falla de presunción en el diseño de los conectores.

QUÉ REVISAR HOY

Cinco minutos en la configuración de tu asistente

La utilidad de este caso para cualquier profesional está en que sugiere una revisión concreta que conviene hacer con cualquier asistente conectado a tus cuentas, con independencia de si usas el producto afectado.

La primera revisión es la **memoria** del asistente, porque ahí se acumulan instrucciones y preferencias que a veces nadie escribió de forma consciente, y la recomendación aplica a cualquier producto de este tipo, ya que una entrada extraña en ese registro puede seguir condicionando las respuestas mucho después de que el episodio que la creó haya sido olvidado.

La segunda revisión son los **conectores**, o sea la lista de servicios a los que le diste acceso alguna vez, y la pregunta útil consiste en verificar si el asistente todavía necesita entrar a tu correo, a tu almacenamiento y a tu calendario, porque cada permiso vigente amplía exactamente aquello que un ataque de este tipo puede alcanzar dentro de tu sesión.

La tercera revisión es de criterio y sirve para lo que viene, porque conviene tratar a cualquier contenido externo que le pidas resumir a un asistente como si pudiera contener instrucciones dirigidas a él, ya que estos sistemas siguen sin distinguir de forma confiable entre lo que deben analizar y lo que deben obedecer, y esa confusión es la raíz técnica de toda esta familia de ataques.

Fuentes. Recomendaciones de Varonis Threat Labs (agosto de 2026) sobre revisión de permisos y detección de actividad inusual. Elaboración propia sobre la aplicación a otros asistentes conectados.

Oscar Obando
CHAVES

> “La inyección sobrevive a los cambios de contraseña, a la revocación de sesión y a la reinscripción del dispositivo, y persiste para siempre.”
>
> — Varonis Threat Labs, divulgación de CoSnitch, 18 de agosto de 2026

Oscar Obando Chaves
oscarobando.ai · linkedin.com/in/oscar-obando-chaves-237495133

Agosto de 2026 · Quito, Ecuador