| | |
|---|---|
| **Alerta ID:** | 099 |
| **Fecha del reporte:** | 12/05/2026 |
| **Entidad:** | Todas las entidades del ecosistema digital |
| **Título:** | Vulnerabilidad crítica “ClaudeBleed” permite secuestro de extensiones de IA en navegadores Chrome |
| **Herramienta de detección** | N/A |
| **Activo involucrado:** | Navegadores Chrome, extensiones de IA, plataformas SaaS, cuentas corporativas, servicios cloud y sistemas integrados con inteligencia artificial |
| **Tipo de alerta:** | Alerta |
| **Nivel de riesgo:** | Alto |

## Objetivo:

Informar a las organizaciones del sector salud en Colombia sobre la identificación de una vulnerabilidad crítica denominada “ClaudeBleed”, la cual afecta la extensión “Claude in Chrome” desarrollada por Anthropic y podría permitir a atacantes secuestrar funcionalidades del asistente de inteligencia artificial, ejecutar acciones privilegiadas y acceder a información sensible sin autorización del usuario.

El propósito de esta alerta es que las entidades del sector salud evalúen inmediatamente el uso de extensiones de inteligencia artificial dentro de navegadores corporativos y validen los riesgos asociados a integraciones entre modelos de lenguaje (LLM), navegadores web, plataformas SaaS y servicios institucionales.

Asimismo, se busca generar conciencia sobre los nuevos riesgos asociados a agentes de IA con capacidades operacionales dentro del navegador, especialmente aquellos con acceso a:

- correos electrónicos;
- archivos corporativos;
- repositorios de código;
- plataformas cloud;
- servicios SaaS;
- sistemas de autenticación;
- y herramientas institucionales integradas.

La creciente adopción de asistentes basados en IA dentro de entornos corporativos está ampliando significativamente la superficie de ataque de las organizaciones, particularmente en sectores críticos como salud, donde existen altos volúmenes de información sensible y múltiples integraciones cloud.

## Descripción:

Investigadores de LayerX identificaron una vulnerabilidad crítica en la extensión “Claude in Chrome”, desarrollada por Anthropic, que permitiría a extensiones maliciosas secuestrar las capacidades del asistente de inteligencia artificial y ejecutar acciones utilizando los privilegios del usuario afectado. La vulnerabilidad fue denominada “ClaudeBleed”.

De acuerdo con los investigadores, la falla se origina por una implementación insegura de límites de confianza (“trust boundaries”) dentro de la extensión. El problema principal radica en que Claude in Chrome confía en cualquier script ejecutado dentro del contexto de claude.ai, sin validar adecuadamente el origen real de las solicitudes enviadas a la extensión.

Esta condición permitiría que una extensión maliciosa instalada en el navegador pueda enviar instrucciones arbitrarias a Claude utilizando mecanismos legítimos de comunicación interna del navegador Chrome. Como resultado, un atacante podría lograr que el asistente de IA:

- acceda a archivos;

- interactúe con correos electrónicos;
- consulte información sensible;
- ejecute acciones automatizadas;
- o exfiltre datos corporativos.

Los investigadores señalaron que incluso extensiones sin permisos especiales podrían explotar esta vulnerabilidad mediante abuso de scripts ejecutados en el contexto principal del navegador. Esto representa un riesgo especialmente importante debido a que rompe parcialmente el modelo tradicional de aislamiento de extensiones implementado por Chrome.

LayerX también indicó que Anthropic publicó una corrección parcial en la versión 1.0.70 de la extensión; sin embargo, los investigadores afirmaron haber logrado evadir las mitigaciones iniciales en aproximadamente tres horas, identificando nuevas formas de explotar el problema mediante el uso de modos privilegiados de ejecución.

El incidente evidencia un problema emergente asociado al rápido crecimiento de herramientas de inteligencia artificial integradas directamente en navegadores y plataformas corporativas. Estas herramientas suelen operar con privilegios elevados y acceso a servicios críticos, incluyendo:

- correo electrónico;
- almacenamiento cloud;
- repositorios de código;
- herramientas colaborativas;
- plataformas SaaS;
- sistemas administrativos;
- y datos sensibles institucionales.

En el contexto del sector salud, este tipo de vulnerabilidades representa un riesgo significativo debido a la utilización creciente de asistentes basados en IA dentro de:

- hospitales;

- universidades médicas;
- laboratorios;
- plataformas clínicas;
- áreas administrativas;
- y entornos académicos.

La explotación exitosa de este tipo de fallas podría derivar en robo de información sensible, compromiso de cuentas institucionales, acceso no autorizado a plataformas cloud y posibles escenarios de movimiento lateral dentro de la infraestructura tecnológica organizacional.

## Consecuencias de la explotación

La explotación de la vulnerabilidad ClaudeBleed puede generar impactos significativos sobre la confidencialidad, integridad y disponibilidad de la información institucional.

Uno de los principales riesgos asociados corresponde al secuestro de funcionalidades del asistente de inteligencia artificial. Debido a que Claude puede interactuar con múltiples servicios y plataformas dentro del navegador, un atacante podría aprovechar la vulnerabilidad para ejecutar acciones utilizando los privilegios legítimos del usuario afectado.

Los investigadores demostraron escenarios donde era posible acceder a archivos almacenados en Google Drive, consultar correos electrónicos, enviar mensajes y exfiltrar información sensible hacia servidores externos sin interacción visible del usuario.

En entornos corporativos y del sector salud, esto podría comprometer:

- información clínica;
- documentos administrativos;
- investigaciones médicas;
- historiales académicos;

- credenciales;
- datos financieros;
- y comunicaciones institucionales.

Otro aspecto crítico corresponde a la posibilidad de abuso de sesiones autenticadas. Dado que las extensiones de IA suelen operar utilizando el contexto autenticado del usuario, un atacante podría heredar permisos y capacidades previamente autorizadas dentro del navegador comprometido.

La vulnerabilidad también incrementa el riesgo de persistencia y movimiento lateral. Una vez comprometida la extensión de IA, los atacantes podrían utilizarla como punto de apoyo para acceder a otros sistemas conectados, plataformas SaaS o servicios cloud integrados dentro de la infraestructura organizacional.

Adicionalmente, este incidente evidencia riesgos emergentes relacionados con agentes de inteligencia artificial capaces de ejecutar acciones automatizadas dentro del navegador.

A diferencia de vulnerabilidades tradicionales, este tipo de ataques combina:

- automatización;
- interacción contextual;
- acceso privilegiado;
- y capacidades operacionales avanzadas.

Esto podría facilitar campañas de espionaje, robo de información y automatización ofensiva mucho más sofisticadas.

## Modo de explotación del ataque

**Fase 1 – Instalación de extensión maliciosa**

El ataque inicia cuando la víctima instala una extensión maliciosa dentro del navegador Chrome. Según los investigadores, la extensión no requiere permisos especiales elevados para ejecutar la explotación.

El atacante puede distribuir la extensión haciéndola pasar por:

- herramientas de productividad;
- utilidades empresariales;
- temas visuales;
- asistentes IA;
- complementos aparentemente legítimos.

Debido a que muchas organizaciones no controlan adecuadamente las extensiones instaladas por los usuarios, este tipo de amenazas puede propagarse fácilmente dentro de entornos corporativos.

---

**Fase 2 – Ejecución de scripts dentro del navegador**

Una vez instalada, la extensión maliciosa ejecuta scripts dentro del contexto del navegador utilizando mecanismos estándar permitidos por Chrome.

Los investigadores identificaron que la extensión Claude in Chrome confía en cualquier script ejecutado dentro del contexto de claude.ai, sin validar adecuadamente el origen real del proceso que envía las instrucciones.

Esta falla permite que extensiones externas interactúen con Claude utilizando canales legítimos de comunicación entre componentes del navegador.

---

**Fase 3 – Secuestro de funcionalidades del asistente de IA**

Posteriormente, el atacante utiliza la extensión maliciosa para enviar instrucciones arbitrarias a Claude.

Debido a la relación de confianza insegura implementada por la extensión vulnerable, Claude interpreta estas instrucciones como legítimas y las ejecuta utilizando los permisos y privilegios del usuario afectado.

Durante esta etapa, el atacante puede:

- solicitar acceso a información;
- automatizar acciones;
- consultar archivos;
- enviar correos;
- interactuar con plataformas SaaS;
- y ejecutar tareas dentro del navegador comprometido.

---

**Fase 4 – Acceso y exfiltración de información sensible**

Una vez obtenido el control operativo sobre el asistente de IA, los atacantes pueden acceder a información sensible disponible dentro del contexto autenticado del usuario.

Los investigadores demostraron escenarios donde era posible:

- consultar correos recientes;
- acceder a archivos en Google Drive;
- extraer información corporativa;
- y transmitir datos hacia infraestructura controlada por el atacante.

En entornos del sector salud, este tipo de acceso podría comprometer:

- información clínica;
- documentos institucionales;
- investigaciones médicas;
- credenciales;
- y datos personales sensibles.

---

**Fase 5 – Persistencia y expansión del compromiso**

Finalmente, el atacante puede utilizar el acceso obtenido para mantener persistencia dentro del entorno afectado y ampliar progresivamente el alcance del compromiso.

Debido a la integración existente entre navegadores, plataformas cloud y servicios SaaS, una explotación exitosa podría facilitar:

- movimiento lateral;
- acceso a servicios corporativos;
- robo de tokens;
- escalamiento de privilegios;
- y expansión hacia otros sistemas institucionales.

Este escenario resulta especialmente crítico en organizaciones que utilizan asistentes de IA con capacidades operacionales amplias dentro del navegador.

## Vulnerabilidades asociadas

**Relaciones de confianza inseguras entre extensiones y asistentes de IA**

- **Descripción:** La vulnerabilidad principal identificada en ClaudeBleed corresponde a una implementación insegura de límites de confianza dentro de la extensión Claude in Chrome. La aplicación confía incorrectamente en cualquier script ejecutado dentro del contexto de claude.ai, sin validar adecuadamente el origen real de las instrucciones.

- **Impacto:** La explotación de esta vulnerabilidad puede permitir que extensiones maliciosas secuestren funcionalidades del asistente de IA y ejecuten acciones utilizando los privilegios legítimos del usuario afectado.

---

**Deficiencias en el modelo de aislamiento de extensiones del navegador**

- **Descripción:** El ataque aprovecha mecanismos legítimos de comunicación entre extensiones y scripts ejecutados dentro del navegador Chrome, debilitando parcialmente el modelo de aislamiento tradicional entre componentes.

- **Impacto:** Los atacantes podrían utilizar extensiones aparentemente legítimas para interactuar con aplicaciones privilegiadas y acceder a información sensible sin generar alertas visibles.

---

**Exposición derivada de asistentes de IA con privilegios elevados**

- **Descripción:** Las extensiones de IA modernas suelen operar con acceso amplio a múltiples servicios corporativos, incluyendo correo electrónico, archivos cloud, plataformas SaaS y herramientas administrativas.

- **Impacto:** Una explotación exitosa puede permitir acceso indirecto a servicios críticos y datos sensibles institucionales, ampliando significativamente el alcance del compromiso.

---

**Riesgos asociados a automatización y agentes de IA**

- **Descripción:** Los asistentes de IA modernos poseen capacidades operacionales avanzadas orientadas a automatizar tareas dentro del navegador y plataformas conectadas.

- **Impacto:** Los atacantes podrían abusar de estas capacidades para automatizar robo de información, exfiltración de datos y ejecución de acciones maliciosas utilizando privilegios legítimos del usuario.

## Recomendaciones de prevención

Las organizaciones del sector salud deben fortalecer sus estrategias preventivas frente al crecimiento acelerado de herramientas de inteligencia artificial integradas en navegadores, plataformas SaaS y entornos corporativos. La incorporación de asistentes basados en IA dentro de procesos administrativos, académicos y clínicos está ampliando significativamente la superficie de ataque organizacional, especialmente cuando estas herramientas poseen acceso a servicios críticos, sesiones autenticadas y plataformas cloud institucionales.

La prevención debe enfocarse no únicamente en la protección del navegador, sino también en el control de integraciones, privilegios, autenticación y acceso a información sensible dentro del ecosistema digital institucional.

**Implementar controles restrictivos sobre extensiones de navegador**

Las extensiones de navegador representan actualmente uno de los principales vectores de exposición dentro de entornos corporativos modernos. Muchas de estas herramientas poseen capacidades avanzadas de interacción con páginas web, sesiones autenticadas, almacenamiento cloud y plataformas SaaS.

Las organizaciones deben establecer políticas institucionales estrictas que regulen la instalación, validación y uso de extensiones dentro de navegadores corporativos, especialmente aquellas relacionadas con asistentes de inteligencia artificial o automatización de tareas.

Resulta fundamental limitar la instalación de complementos no autorizados y validar cuidadosamente los permisos solicitados por cada extensión antes de permitir su uso institucional.

**Controles recomendados:**

- Implementar listas blancas de extensiones autorizadas.
- Restringir instalación libre de complementos.
- Bloquear extensiones de origen desconocido.
- Aplicar administración centralizada de navegadores.
- Validar permisos solicitados por extensiones.
- Deshabilitar complementos innecesarios.
- Monitorear nuevas instalaciones y cambios de configuración.

**Limitar privilegios de asistentes de inteligencia artificial**

Los asistentes basados en IA suelen integrarse con múltiples servicios institucionales, incluyendo correo electrónico, almacenamiento cloud, plataformas colaborativas y aplicaciones SaaS. Cuando estos asistentes operan con privilegios excesivos, una explotación exitosa puede ampliar significativamente el impacto de un compromiso.

Las organizaciones deben aplicar el principio de mínimos privilegios y restringir el acceso automático de herramientas IA a información crítica o plataformas sensibles.

**Controles recomendados:**

- Restringir acceso automático a servicios críticos.
- Limitar integración con plataformas administrativas.
- Segmentar accesos según perfiles de usuario.
- Validar integraciones SaaS antes de habilitarlas.
- Implementar aprobaciones explícitas para acciones sensibles.
- Revisar periódicamente permisos otorgados a asistentes IA.
- Restringir acceso a información clínica y financiera.

---

**Fortalecer la seguridad de navegadores corporativos**

El navegador se ha convertido en un punto central de interacción entre usuarios, plataformas cloud y servicios corporativos. Esto convierte a los navegadores en objetivos prioritarios para actores maliciosos que buscan comprometer sesiones autenticadas y abusar de integraciones privilegiadas.

Las entidades deben fortalecer las configuraciones de seguridad del navegador y reducir la exposición frente a ejecución de scripts maliciosos, abuso de extensiones y secuestro de sesiones.

**Controles recomendados:**

- Mantener navegadores permanentemente actualizados.
- Aplicar hardening sobre configuraciones del navegador.
- Restringir ejecución de scripts no confiables.
- Configurar políticas seguras de cookies y sesiones.
- Implementar aislamiento de navegación cuando sea posible.
- Bloquear acceso a sitios maliciosos conocidos.
- Deshabilitar funciones innecesarias del navegador.

---

**Implementar estrategias Zero Trust y protección de identidad**

Las amenazas modernas asociadas a inteligencia artificial y automatización ofensiva requieren mecanismos de validación continua de identidad y contexto. El modelo Zero Trust resulta especialmente importante debido a que muchas herramientas IA operan utilizando sesiones autenticadas legítimas.

Las organizaciones deben asumir que ninguna sesión, usuario o dispositivo debe considerarse confiable por defecto.

**Controles recomendados:**

- Implementar MFA resistente a phishing.
- Aplicar autenticación adaptativa.
- Validar contexto y ubicación de acceso.
- Restringir accesos privilegiados.
- Monitorear comportamiento anómalo de sesiones.
- Implementar PAM para cuentas administrativas.
- Aplicar segmentación y microsegmentación de accesos.

**Establecer políticas corporativas para uso de inteligencia artificial**

Muchas organizaciones han adoptado herramientas basadas en IA sin definir controles claros relacionados con privacidad, permisos, tratamiento de información sensible y riesgos de integración tecnológica.

Es fundamental establecer lineamientos institucionales que regulen el uso seguro de asistentes de inteligencia artificial dentro de entornos corporativos y clínicos.

**Controles recomendados:**

- Definir políticas institucionales de uso de IA.
- Restringir carga de información sensible en asistentes IA.
- Validar herramientas IA antes de su adopción.
- Implementar procesos de evaluación de riesgos tecnológicos.
- Capacitar usuarios sobre riesgos asociados a IA.
- Establecer controles sobre integraciones cloud.
- Definir lineamientos para tratamiento de datos sensibles.

## Recomendaciones de detección y mitigación

Las organizaciones deben fortalecer sus capacidades de monitoreo y detección frente a amenazas asociadas al abuso de extensiones, asistentes de inteligencia artificial y automatización maliciosa dentro del navegador.

Debido a que este tipo de amenazas puede utilizar sesiones legítimas y comportamientos aparentemente válidos, resulta fundamental implementar capacidades avanzadas de análisis contextual y monitoreo basado en comportamiento.

---

**Monitorear actividad anómala relacionada con extensiones y asistentes IA**

Las organizaciones deben implementar monitoreo continuo sobre extensiones instaladas y comportamiento de asistentes IA dentro de navegadores corporativos.

Es importante identificar:

- ejecución anómala de scripts;
- automatización sospechosa;
- nuevas extensiones;
- conexiones no habituales;
- y actividades asociadas a exfiltración de información.

**Controles recomendados:**

- Monitorear instalación de extensiones.
- Detectar cambios no autorizados en navegadores.
- Generar alertas sobre automatización sospechosa.
- Integrar logs de navegador en SIEM.
- Detectar conexiones anómalas hacia servicios externos.
- Monitorear actividad asociada a asistentes IA.
- Correlacionar eventos cloud y SaaS.

---

**Implementar capacidades avanzadas de monitoreo y análisis de comportamiento**

Las amenazas modernas potenciadas por automatización e inteligencia artificial pueden modificar dinámicamente sus patrones de comportamiento y dificultar la detección basada únicamente en firmas tradicionales.

Las organizaciones deben fortalecer capacidades de monitoreo contextual y análisis de comportamiento orientadas a detectar anomalías relacionadas con autenticaciones, acceso a información sensible y automatización ofensiva.

**Controles recomendados:**

- Implementar SIEM con correlación avanzada.
- Integrar capacidades UEBA.
- Implementar soluciones XDR/EDR.
- Detectar accesos geográficos anómalos.
- Monitorear creación de tokens y sesiones.
- Analizar comportamiento privilegiado.
- Detectar movimientos laterales sospechosos.

---

**Fortalecer monitoreo cloud y protección de información sensible**

Debido a que muchos asistentes IA interactúan directamente con plataformas cloud y servicios SaaS, las organizaciones deben fortalecer monitoreo sobre accesos, transferencias y comportamiento relacionado con información crítica.

Resulta especialmente importante monitorear plataformas:

- Microsoft 365;
- Google Workspace;
- almacenamiento cloud;
- y servicios colaborativos.

**Controles recomendados:**

- Implementar DLP (Data Loss Prevention).

- Monitorear transferencias masivas de información.
- Detectar accesos inusuales a archivos sensibles.
- Auditar actividad sobre plataformas cloud.
- Configurar alertas por exfiltración de datos.
- Monitorear cambios de permisos y compartición.
- Analizar comportamiento anómalo sobre SaaS.

---

**Implementar gestión acelerada de vulnerabilidades**

Las vulnerabilidades asociadas a asistentes IA y extensiones pueden evolucionar rápidamente y ser explotadas en cortos periodos de tiempo. Las organizaciones deben acelerar sus procesos de identificación, evaluación y remediación de vulnerabilidades.

**Controles recomendados:**

- Mantener extensiones y navegadores actualizados.
- Monitorear boletines de seguridad.
- Realizar revisiones periódicas de configuraciones.
- Validar integraciones inseguras.
- Ejecutar pruebas de seguridad continuas.
- Implementar escaneos de vulnerabilidades.
- Priorizar activos expuestos a Internet.

## Recomendaciones de respuesta

En caso de sospecha de compromiso relacionado con asistentes de inteligencia artificial o extensiones maliciosas, las organizaciones deben activar inmediatamente sus procedimientos de respuesta a incidentes y ejecutar acciones orientadas a contener, erradicar y recuperar los sistemas afectados.

La respuesta debe enfocarse especialmente en la protección de sesiones autenticadas, credenciales, plataformas cloud y accesos privilegiados.

---

**Contener inmediatamente sistemas potencialmente comprometidos**

Ante cualquier indicio de explotación, resulta fundamental aislar los equipos afectados y deshabilitar extensiones sospechosas para evitar expansión del compromiso y exfiltración adicional de información.

**Controles recomendados:**

- Deshabilitar extensiones sospechosas.
- Aislar equipos comprometidos.
- Revocar sesiones activas.
- Invalidar tokens de autenticación.
- Bloquear conexiones sospechosas.
- Restringir accesos privilegiados temporalmente.
- Segmentar sistemas afectados.

---

**Ejecutar análisis forense y validación de integridad**

Las organizaciones deben realizar análisis orientados a identificar:

- scripts maliciosos;
- actividad automatizada;
- exfiltración de información;
- abuso de sesiones;
- y persistencia dentro del navegador o plataformas cloud.

**Controles recomendados:**

- Revisar logs de autenticación y navegación.
- Analizar actividad cloud y SaaS.
- Verificar integridad de configuraciones.
- Identificar extensiones maliciosas.
- Detectar accesos no autorizados.
- Revisar actividad administrativa reciente.
- Ejecutar análisis forense sobre endpoints afectados.

---

**Proteger credenciales y accesos institucionales**

Debido a que este tipo de ataques puede involucrar abuso de sesiones autenticadas y robo de tokens, resulta fundamental proteger inmediatamente credenciales institucionales y accesos privilegiados.

**Controles recomendados:**

- Rotar contraseñas comprometidas.
- Revocar tokens y sesiones activas.
- Restablecer credenciales administrativas.
- Revisar cuentas privilegiadas.
- Implementar MFA reforzado.
- Monitorear reutilización de credenciales.
- Auditar accesos recientes.

---

**Restaurar operación segura y fortalecer controles posteriores al incidente**

Una vez contenido el incidente, las organizaciones deben validar la integridad de los sistemas afectados y fortalecer controles para prevenir nuevos compromisos relacionados con asistentes IA o extensiones vulnerables.

**Controles recomendados:**

- Restaurar configuraciones seguras.
- Validar integridad de navegadores y extensiones.
- Actualizar políticas de seguridad.
- Reforzar monitoreo posterior al incidente.
- Implementar lecciones aprendidas.
- Actualizar playbooks SOC/CSIRT.
- Capacitar usuarios y equipos técnicos.

## Conclusiones

La vulnerabilidad ClaudeBleed evidencia los riesgos emergentes asociados a la rápida adopción de asistentes de inteligencia artificial integrados directamente dentro de navegadores y plataformas corporativas.

La combinación entre automatización, privilegios elevados y relaciones de confianza inseguras puede permitir que actores maliciosos secuestren funcionalidades de agentes de IA y utilicen capacidades legítimas del usuario para acceder a información sensible y ejecutar acciones no autorizadas.

Este incidente demuestra que las extensiones basadas en inteligencia artificial están ampliando significativamente la superficie de ataque organizacional, especialmente en entornos donde los navegadores funcionan como punto central de acceso a plataformas SaaS, servicios cloud y herramientas institucionales.

Las organizaciones del sector salud deben fortalecer urgentemente controles relacionados con:

- seguridad de navegadores;
- extensiones corporativas;
- asistentes de IA;
- monitoreo cloud;
- y gestión de accesos privilegiados.

El crecimiento acelerado de herramientas de IA con capacidades operacionales avanzadas continuará generando nuevos escenarios de riesgo que requerirán estrategias de seguridad más robustas, monitoreo continuo y controles específicos para entornos basados en inteligencia artificial.

## Fuentes:


- https://www.securityweek.com/vulnerability-in-claude-extension-for-chrome-exposes-ai-agent-to-takeover/?utm_source=chatgpt.com
- https://cybernews.com/security/claude-code-chrome-extension-flaw-fix-hacked/?utm_source=chatgpt.com
- https://www.csoonline.com/article/4168867/claude-in-chrome-is-taking-orders-from-the-wrong-extensions.html?utm_source=chatgpt.com
- https://hackread.com/claudebleed-vulnerability-hackers-claude-chrome-extension/?utm_source=chatgpt.com
- https://www.koi.ai/blog/shadowprompt-how-any-website-could-have-hijacked-anthropic-claude-chrome-extension?utm_source=chatgpt.com


En caso de dudas, inquietudes o requerir asistencia adicional relacionada con esta alerta de seguridad, puede comunicarse directamente con el **CSIRT Salud** a través de las líneas telefónicas **(+57) 316 893 1490 - 318 155 3570** o mediante el correo electrónico **csirtsalud@minsalud.gov.co**. Nuestro equipo está disponible para brindar el acompañamiento necesario.