Google anunció este lunes un conjunto de nuevas funciones de seguridad en Chrome, tras la incorporación de capacidades de inteligencia artificial (IA) agente al navegador web.

Con ello, la compañía afirmó que ha incorporado defensas en capas para dificultar que actores maliciosos exploten inyecciones de indicaciones indirectas derivadas de la exposición a contenido web no confiable y provoquen daños.

La característica principal es un *User Alignment Critic*, un sistema que emplea un segundo modelo para evaluar de forma independiente las acciones del agente, aislándolo de instrucciones maliciosas. Este enfoque complementa técnicas ya existentes de Google, como *spotlighting*, que obliga al modelo a seguir las instrucciones del usuario y del sistema, en lugar de acatar lo incrustado en una página web.

"El User Alignment Critic se ejecuta después de que finaliza la planificación para verificar nuevamente cada acción propuesta", dijo Google. *"Su objetivo principal es comprobar la alineación con la tarea: determinar si la acción planteada cumple con el objetivo expresado por el usuario. Si no está alineada, el Alignment Critic la rechazará."*

Este componente solo visualiza metadatos sobre la acción propuesta y se evita que acceda a contenido web no confiable, lo cual le impide ser manipulado por instrucciones maliciosas incluidas en algún sitio. Con el User Alignment Critic, la intención es añadir protecciones frente a intentos de extraer datos o desviar los objetivos legítimos hacia acciones buscadas por un atacante.

"Cuando una acción es rechazada, el Critic ofrece retroalimentación al modelo de planificación para que reformule su plan, y el planificador puede devolver el control al usuario si ocurren fallos repetidos", señaló Nathan Parker del equipo de seguridad de Chrome.

Google también está imponiendo un mecanismo denominado *Agent Origin Sets* para garantizar que el agente solo acceda a datos provenientes de orígenes relevantes para la tarea o que el usuario haya decidido compartir con él. Esto busca impedir la evasión del

aislamiento entre sitios, en la cual un agente comprometido pudiera interactuar con dominios arbitrarios y extraer información de sitios donde el usuario tiene sesión iniciada.

Esto se implementa mediante una función de filtrado que define cuáles orígenes están relacionados con la actividad y los separa en dos grupos:

- Orígenes de solo lectura, desde los cuales el modelo Gemini de Google puede consumir contenido.
- Orígenes de lectura y escritura, a los cuales el agente puede escribir o hacer clic además de leer.

"Esta separación garantiza que solo datos de un conjunto limitado de orígenes estén disponibles para el agente, y que esta información únicamente se transfiera hacia los orígenes que permiten escritura", explicó Google. *"De esta forma se acota el vector de riesgo referente a fugas de datos entre orígenes."*

Al igual que con el User Alignment Critic, la función de filtrado no está expuesta a contenido web no confiable. Además, el planificador debe obtener su aprobación antes de añadir nuevos orígenes, aunque sí puede utilizar contexto de páginas que el usuario haya compartido explícitamente durante una sesión.

Otro pilar fundamental de la nueva arquitectura de seguridad se relaciona con proporcionar mayor transparencia y control al usuario, permitiendo que el agente genere un registro de trabajo para su revisión y solicite aprobación explícita antes de navegar hacia sitios sensibles, como portales bancarios o médicos, autorizar inicios de sesión mediante Google Password Manager o realizar acciones web como compras, pagos o envío de mensajes.

Finalmente, el agente revisa cada página en busca de inyecciones de indicaciones indirectas y trabaja junto con Safe Browsing y el sistema local de detección de estafas para bloquear contenido potencialmente sospechoso.

"Este clasificador de inyección de indicaciones se ejecuta en paralelo al proceso de inferencia

del modelo de planificación, y evitará que se tomen acciones basadas en contenido que el clasificador determine que intenta manipular al modelo para hacer algo no alineado con la meta del usuario", indicó Google.

Para impulsar la investigación y detectar fallos en el sistema, la compañía señaló que pagará hasta $20,000 por demostraciones que logren vulnerar los límites de seguridad. Esto incluye inyecciones indirectas que permitan a un atacante:

- Ejecutar acciones no autorizadas sin confirmación.
- Extraer información sensible sin oportunidad efectiva de aprobación por parte del usuario.
- Saltarse una mitigación que debería haber impedido el ataque.

"Al extender principios fundamentales como el aislamiento por origen y las defensas en capas, y al introducir una arquitectura de modelos confiables, estamos construyendo una base segura para las experiencias agente de Gemini en Chrome", señaló Google. *"Seguiremos comprometidos con la innovación continua y la colaboración con la comunidad de seguridad para garantizar que los usuarios de Chrome exploren esta nueva era de la web de manera segura."*

El anuncio llega tras una investigación de Gartner, que recomendó a las empresas bloquear el uso de navegadores con IA agente hasta que los riesgos asociados —como inyecciones indirectas, acciones erróneas del agente y pérdida de datos— puedan ser gestionados adecuadamente.

La investigación también advierte de un posible escenario en el que los empleados *"podrían verse tentados a usar navegadores con IA y automatizar tareas obligatorias, repetitivas y poco interesantes"*. Esto incluiría casos en los que alguien evite entrenamientos de ciberseguridad obligatorios indicando al navegador con IA que los complete en su lugar.

"Los navegadores agentes, o lo que muchos llaman navegadores con IA, tienen el potencial de transformar cómo los usuarios interactúan con sitios web y automatizan transacciones,

pero también introducen riesgos críticos de ciberseguridad”, indicó la firma consultora. *“Los CISOs deben bloquear todos los navegadores con IA por el futuro cercano para minimizar la exposición al riesgo.”*

El desarrollo se produce mientras el Centro Nacional de Ciberseguridad (NCSC) del Reino Unido afirmó que los modelos de lenguaje grandes (LLMs) pueden presentar una clase persistente de vulnerabilidad conocida como inyección de indicaciones y que este problema no puede eliminarse por completo.

“Los modelos de lenguaje grandes (LLMs) actuales simplemente no imponen un límite de seguridad entre instrucciones y datos dentro de una indicación”, explicó David C., director técnico de Investigación de Plataformas del NCSC. *“Por ello, las protecciones deben centrarse más en salvaguardas deterministas (no basadas en LLM) que restrinjan las acciones del sistema, en lugar de intentar únicamente impedir que contenido malicioso llegue al LLM.”*