La división DeepMind de Google anunció este lunes un agente impulsado por inteligencia artificial (IA) llamado CodeMender, diseñado para detectar automáticamente vulnerabilidades en el código, corregirlas y reescribirlas, con el fin de evitar posibles explotaciones futuras.

Este proyecto se suma a los esfuerzos continuos de la compañía por mejorar la detección de vulnerabilidades mediante IA, como lo ha hecho anteriormente con iniciativas como Big Sleep y OSS-Fuzz.

Según DeepMind, este agente ha sido creado para actuar de manera tanto reactiva como proactiva: repara fallos de seguridad tan pronto como se identifican y también reescribe bases de código ya existentes con el objetivo de eliminar tipos completos de vulnerabilidades.

"Al generar y aplicar parches de seguridad de alta calidad de forma automática, el agente con IA de CodeMender permite a los desarrolladores y mantenedores concentrarse en lo que mejor saben hacer: crear buen software", afirmaron los investigadores de DeepMind, *Raluca Ada Popa y Four Flynn*.

"Durante los últimos seis meses en los que hemos trabajado en CodeMender, ya hemos contribuido con 72 correcciones de seguridad a proyectos de código abierto, incluyendo algunos con hasta 4,5 millones de líneas de código", añadieron.

En su funcionamiento interno, CodeMender utiliza los modelos Gemini Deep Think de Google para identificar, depurar y resolver vulnerabilidades de seguridad abordando directamente su causa raíz, y validando que las soluciones no introduzcan errores o fallos secundarios.

Además, según Google, el agente también incorpora una herramienta de crítica basada en modelos de lenguaje de gran escala (LLM) que resalta las diferencias entre el código original y el modificado, para comprobar que los cambios no generen regresiones, y, si es necesario, los corrige por sí mismo.

Google también señaló que planea contactar progresivamente a los mantenedores de

proyectos de código abierto críticos para ofrecerles parches generados por CodeMender, y solicitar su retroalimentación, con el objetivo de utilizar esta herramienta para mantener sus bases de código seguras.

Este desarrollo ocurre en paralelo a la creación del Programa de Recompensas por Vulnerabilidades de IA (AI VRP), a través del cual se pueden reportar fallos relacionados con IA en los productos de Google —como inyecciones de prompts, *jailbreaks* y problemas de alineación—, y recibir recompensas que pueden alcanzar hasta los $30,000 dólares.

En junio de 2025, Anthropic reveló que modelos de varios desarrolladores adoptaron comportamientos maliciosos tipo "infiltrado interno" cuando eso era la única vía para evitar ser reemplazados o cumplir sus objetivos, y que los modelos de lenguaje *"se comportaban menos mal cuando sabían que estaban en una prueba, y actuaban peor cuando pensaban que la situación era real"*.

Dicho esto, la generación de contenido que infringe políticas, la evasión de filtros de seguridad, las alucinaciones, errores fácticos, extracción de instrucciones del sistema y problemas de propiedad intelectual no están cubiertos por el AI VRP.

Google, que ya había formado un equipo especializado en seguridad de IA (AI Red Team) como parte de su Marco de IA Segura (SAIF), ha lanzado también una segunda versión de dicho marco, centrada en riesgos asociados a agentes autónomos, como filtración de datos o acciones no deseadas, así como los controles necesarios para mitigarlos.

La compañía también destacó su compromiso con el uso de la inteligencia artificial para fortalecer la seguridad y la protección digital, y emplear esta tecnología para dar ventaja a los defensores ante amenazas crecientes por parte de ciberdelincuentes, estafadores y atacantes respaldados por estados.