Google ha comunicado su ampliación del Programa de Recompensas por Vulnerabilidades (VRP) con el propósito de remunerar a los investigadores que identifiquen escenarios de ataque diseñados para sistemas de inteligencia artificial generativa (IA), en un esfuerzo por reforzar la seguridad y fiabilidad de la IA.

> Laurie Richardson y Royal Hansen de Google han señalado que *«la IA generativa plantea preocupaciones distintas y novedosas en comparación con la seguridad digital tradicional, como la posibilidad de sesgos injustos, manipulación de modelos o interpretaciones erróneas de datos (alucinaciones)».*

Entre las categorías incluidas en el ámbito del programa se encuentran las inyecciones de instrucciones, la filtración de datos confidenciales de conjuntos de datos de entrenamiento, la manipulación de modelos, los ataques de perturbación adversaria que inducen a errores en la clasificación y el robo de modelos.

Es importante mencionar que Google, a principios de julio, estableció un Equipo Rojo de IA para abordar las amenazas a los sistemas de IA como parte de su Marco de IA Segura (SAIF).

Además, como parte de su compromiso con la IA segura, Google también ha anunciado esfuerzos para fortalecer la cadena de suministro de IA a través de iniciativas de seguridad de código abierto existentes, como Niveles de la Cadena de Suministro para Artefactos de Software (SLSA) y Sigstore.

> *«Las firmas digitales, como las que ofrece Sigstore, permiten a los usuarios verificar que el software no ha sido alterado o reemplazado»,* destacó Google.

> *«Los metadatos, como la procedencia de SLSA, proporcionan información sobre el contenido del software y cómo se construyó, lo que permite a los consumidores*

> *garantizar la compatibilidad de licencias, identificar vulnerabilidades conocidas y detectar amenazas más avanzadas».*

Este anuncio coincide con la presentación por parte de OpenAI de un nuevo equipo interno de Preparación que se encargará de «*rastrear, evaluar, prever y proteger*» contra riesgos catastróficos relacionados con la IA generativa, abarcando amenazas en ciberseguridad, químicas, biológicas, radiológicas y nucleares (CBRN).

Además, ambas empresas, junto con Anthropic y Microsoft, han anunciado la creación de un Fondo de Seguridad de IA de $10 millones, enfocado en promover la investigación en el ámbito de la seguridad de la IA.