Google

Marco para la IA segura de Google (SAIF, por sus siglas en inglés)

Ante el rápido avance de la IA, es importante contar con estrategias de gestión de riesgos eficaces que evolucionen con ella. Para ayudar a conseguirlo, hemos lanzado el Marco para la IA Segura de Google, un marco de trabajo conceptual para proteger los sistemas de IA. SAIF tiene seis objetivos fundamentales:

1. Extender nuestros sólidos cimientos de seguridad a todo el ecosistema de la IA

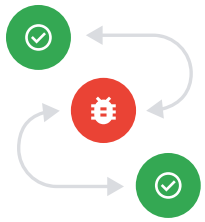
Aprovechar una infraestructura segura por defecto y el conocimiento especializado adquirido durante las últimas dos décadas para proteger los sistemas, las aplicaciones y los usuarios del ecosistema de la IA. Al mismo tiempo, facilitar el conocimiento técnico a las empresas, que les permita mantenerse al día de los avances en IA y comenzar a escalar y adaptar las medidas de seguridad de su infraestructura en el contexto de la IA y de los modelos de amenazas en evolución. Por ejemplo, en el caso de determinadas técnicas de inyección que existen desde hace tiempo, como la inyección de SQL, las empresas pueden adaptar sus medidas de mitigación, como el saneamiento y la limitación de entradas, para mejorar la defensa frente a ataques por inyección de prompts.

2. Ampliar las capacidades de detección y respuesta para llevar la IA a los sistemas de defensa antiamenazas de las organizaciones

Detectar incidentes de ciberseguridad asociados con la IA y responder a ellos con rapidez ampliando la inteligencia de amenazas y otras capacidades. Para las empresas, esto implica supervisar las entradas y los resultados de los sistemas de IA generativa para detectar anomalías y usar inteligencia de amenazas para adelantarse a los ataques. Esta tarea suele requerir la colaboración de los equipos de confianza y seguridad, inteligencia de amenazas y prevención de usos indebidos.

3. Automatizar los sistemas de defensa para mantenerlos al día de las amenazas existentes y de las nuevas

Recurrir a las últimas innovaciones en el ámbito de la IA para mejorar la escala y la velocidad de respuesta a los incidentes de seguridad. Puesto que es probable que muchos atacantes utilicen la IA para ampliar su impacto, es importante aprovechar las capacidades de esta tecnología, tanto las nuevas, como las existentes, para responder y protegerse frente a sus ataques con rapidez y por un coste razonable.

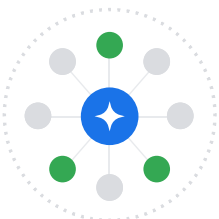


4. Unificar los controles a nivel de plataforma para garantizar una seguridad uniforme en toda la organización

Alinear los marcos de control para facilitar la mitigación de los riesgos derivados de la IA, y escalar las medidas de protección en diferentes plataformas y herramientas. De esta forma, se garantiza que todas las aplicaciones de IA cuenten con una protección óptima que sea escalable y rentable. En Google, esto incluye extender nuestras medidas de protección intrínsecamente seguras a las plataformas de IA como Vertex AI y Security AI Workbench, además de integrar controles y defensas en el ciclo de vida del desarrollo de software. Las soluciones que abordan casos de uso generales, como la API de Perspective, pueden contribuir a que toda la organización se beneficie de las medidas de seguridad más avanzadas.

5. Adaptar controles para ajustar las medidas de mitigación y crear procesos de respuesta más rápidos para la implementación de la IA

Verificar constantemente las implementaciones mediante el aprendizaje continuo, y desarrollar medidas de detección y protección para hacer frente a un entorno de amenazas cambiante. Esto requiere usar técnicas como el aprendizaje por refuerzo basado en incidentes y respuesta de los usuarios, e implica realizar acciones como actualizar los conjuntos de datos de entrenamiento, afinar los modelos para responder estratégicamente a los ataques y permitir que el software con el que se construyen los modelos incorpore funciones de seguridad contextual adicionales (p. ej., la detección de comportamientos anómalos). Las organizaciones también pueden realizar regularmente ejercicios de Red Team para mejorar los procesos de control de la seguridad en los productos y funciones basados en IA.



6. Contextualizar los riesgos del sistema de IA en los procesos empresariales de su entorno

Realizar evaluaciones de riesgos en extremo a extremo de los procesos de implementación de la IA en las organizaciones. Aquí se incluye realizar una evaluación de los riesgos empresariales de extremo a extremo como el linaje de datos, la validación y la supervisión del comportamiento operativo de determinados tipos de aplicaciones. Además, las organizaciones deben implantar verificaciones automatizadas que validen el funcionamiento de la IA.