Google

Marco Seguro de IA de Google

La tecnología de IA avanza rápidamente y es importante que las estrategias de gestión de riesgos evolucionen de manera paralela. Para lograr que dicha evolución sea un éxito, presentamos el Marco Seguro de IA (SAIF, por su sigla en inglés), un marco conceptual para crear sistemas de IA seguros. SAIF cuenta con seis elementos principales:

1. Desarrollar cimientos sólidos de seguridad para el ecosistema de IA

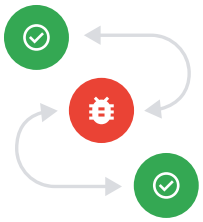
Implementar las protecciones de infraestructura seguras por defecto junto a la experiencia que se ha adquirido a lo largo de las últimas dos décadas con el fin de proteger los sistemas de IA, sus aplicaciones y usuarios. Simultáneamente, desarrollar una pericia organizacional para seguir el ritmo de los avances de la IA y aumentar y adaptar la infraestructura de protección relativa a la IA y los crecientes modelos de amenazas. Por ejemplo, ciertas técnicas de inyección -como la inyección SQL- han existido desde hace tiempo y las organizaciones pueden adaptar mitigaciones, como la sanitización y limitación de entrada, para defenderse mejor contra los ataques repentinos estilo inyección.

2. Incrementar tanto la detección como la respuesta para incluir la IA en el universo de amenazas de una organización

Detectar y responder de manera oportuna a los incidentes cibernéticos relacionados con la IA ampliando los programas de inteligencia contra amenazas y otras capacidades. En cuanto a las organizaciones, esto incluye la supervisión de entradas y salidas de sistemas de IA generativa para detectar anomalías y utilizar programas de inteligencia contra amenazas con el fin de anticipar cualquier ataque. Este esfuerzo generalmente requiere de la colaboración de equipos de confianza y seguridad, programas de inteligencia contra amenazas, así como equipos de salvaguardias contra el abuso.

3. Automatizar sistemas de defensa para avanzar al mismo ritmo que las amenazas actuales y las nuevas

Emplear las innovaciones más recientes de la IA para mejorar tanto la escala como la velocidad de los esfuerzos de respuesta ante cualquier incidente de seguridad. Es probable que los adversarios utilicen la IA para escalar su impacto, por lo que es importante utilizar la IA y sus capacidades actuales y emergentes para protegernos de dichas amenazas de manera diestra y rentable.

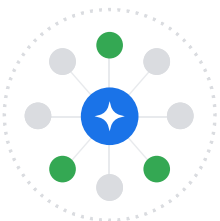


4. Unificar los controles a nivel de plataforma para garantizar una seguridad constante en toda la organización

Alinear los marcos de control para apoyar la mitigación de riesgos de IA e implementar protecciones a través de diferentes plataformas y herramientas para asegurar que las mejores protecciones estén disponibles en todas las aplicaciones de IA de una manera escalable y rentable. En Google, esto implica ampliar las protecciones de seguridad por defecto a las plataformas de IA, como Vertex AI y Security AI Workbench, y crear controles y protecciones dentro del ciclo de vida del desarrollo de software. Las aplicaciones que abordan los casos de uso generales, como Perspective API, pueden ayudar a que la organización completa se beneficie de protecciones de última generación.

5. Adaptar los controles para ajustar las mitigaciones y crear ciclos de retroalimentación más rápidos para el despliegue de la IA

Probar las implementaciones de manera constante mediante un aprendizaje continuo y desarrollar detecciones y protecciones para abordar un entorno de amenazas en constante cambio. Esto incluye diferentes técnicas, como el aprendizaje por refuerzo basado en incidentes y la retroalimentación del usuario, e involucra pasos como actualizar los conjuntos de datos de entrenamiento, refinar los modelos para responder estratégicamente a los ataques y permitir que el software que se utiliza para compilar modelos integre una mayor seguridad en contexto (p. ej., detección de comportamiento anómalo). Además, las organizaciones pueden realizar ejercicios de simulacros de ataque con regularidad para mejorar la garantía de seguridad de las capacidades y los productos de IA.



6. Contextualizar los riesgos de los sistemas de IA en los procesos del entorno de la empresa

Realizar evaluaciones de riesgo integrales con respecto al modo en que las organizaciones desplegarán la IA. Aquí se incluye la evaluación del riesgo empresarial integral, como el linaje y validación de datos y el control del comportamiento operativo de determinados tipos de aplicaciones. Además, las organizaciones deben crear controles automáticos para validar el desempeño de la IA.