Google

Secure AI Framework di Google

L'IA sta avanzando rapidamente ed è importante che con essa si evolvano anche delle strategie efficaci per la gestione dei rischi. Per contribuire a questa evoluzione abbiamo creato il Secure AI Framework (SAIF), un modello concettuale per rendere più sicuri i sistemi d'intelligenza artificiale. SAIF si basa su sei elementi di base:

1. Costruire delle basi di sicurezza solide per l'ecosistema IA

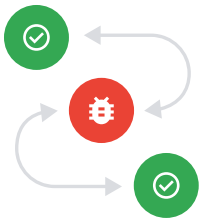
Sfruttare le protezioni dell'infrastruttura secure-by-default e le competenze sviluppate negli ultimi vent'anni per proteggere i sistemi IA, le applicazioni e gli utenti. Allo stesso tempo, sviluppare competenze organizzative per tenere il passo con i progressi dell'IA e iniziare a scalare e adattare le protezioni dell'infrastruttura nel contesto dell'IA e dei modelli di minaccia in continua evoluzione. Per esempio, gli attacchi informatici come l'SQL injection esistono da tempo e le aziende possono adeguare le azioni di mitigazione, come l'input sanitization e il limiting, per contribuire a proteggersi meglio dagli attacchi di prompt injection.

2. Estendere il rilevamento e la risposta per portare l'IA nell'universo delle minacce informatiche delle aziende

Rilevare e rispondere per tempo agli incidenti di sicurezza informatica correlati all'IA ampliando la threat intelligence e altre capacità. Per le aziende significa anche monitorare gli input e gli output dei sistemi di IA generativa per rilevare anomalie e utilizzare la threat intelligence per prevenire gli attacchi. Questo sforzo richiede solitamente la collaborazione tra i team di trust & safety, threat intelligence e contrasto agli abusi.

3. Automatizzare le difese per tenere il passo con le nuove minacce e con quelle esistenti

Sfruttare le ultime innovazioni nel campo dell'IA per migliorare la portata e la velocità della risposta agli incidenti di sicurezza. Gli avversari utilizzeranno probabilmente l'IA per ridimensionare il proprio impatto per cui è importante avvalersi dell'IA e delle sue capacità emergenti per restare agili e garantire una protezione efficiente anche sotto il profilo dei costi.

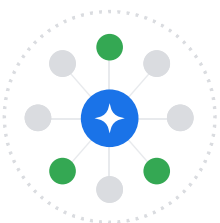


4. Armonizzare i controlli a livello di piattaforma per assicurare una sicurezza omogenea in tutta l'azienda

Allineare i framework di controllo per sostenere la mitigazione dei rischi legati all'IA e adattare le protezioni su diverse piattaforme e strumenti per garantire le migliori protezioni possibili per tutte le applicazioni IA in maniera scalabile ed efficiente in termini di costo. Per Google questo prevede anche di estendere le protezioni secure-by-default alle piattaforme IA come Vertex AI e Security AI Workbench, nonché integrare controlli e protezioni nel ciclo di sviluppo del software. Le funzionalità che affrontano casi di utilizzo generali, come Perspective API, possono aiutare l'intera azienda a trarre beneficio dalle protezioni avanzate.

5. Adattare i controlli per regolare le mitigazioni e creare cicli di feedback più veloci per l'implementazione dell'IA

Testare costantemente le implementazioni attraverso l'apprendimento continuo ed evolvere il rilevamento e le protezioni per adeguarsi ai cambiamenti che interessano l'ambiente delle minacce. Questa operazione include tecniche come l'apprendimento rinforzato basato su incidenti e feedback degli utenti, e prevede passaggi specifici come l'aggiornamento dei dataset di addestramento, la messa a punto dei modelli per rispondere agli attacchi in modo strategico e la predisposizione del software utilizzato affinché costruisca dei modelli per integrare un ulteriore livello di sicurezza nel contesto (ad es. rilevare comportamenti anomali). Le aziende possono anche condurre delle regolari esercitazioni Red Team per migliorare la garanzia di sicurezza per i prodotti e le capacità basati sull'IA.



6. Contestualizzare i rischi dei sistemi IA nei processi aziendali circostanti

Condurre valutazioni dei rischi end-to-end in relazione a come le aziende implementeranno l'IA. Ciò include una valutazione del rischio aziendale end-to-end, come il data lineage, la convalida e il monitoraggio del comportamento operativo per certi tipi di applicazioni. Inoltre, le aziende dovrebbero svolgere dei controlli automatici per convalidare le prestazioni dell'IA.