Google

Le framework d'IA sécurisé de Google

L'IA progresse à pas de géants. Il est donc essentiel que les stratégies de gestion du risque efficaces suivent ce rythme. Pour y parvenir, nous avons créé notre framework d'IA sécurisé (SAIF), un framework conceptuel visant à renforcer la sécurité des systèmes d'IA. Le SAIF s'articule autour de six principes :

1. Développer des bases solides pour la sécurité de l'écosystème de l'IA

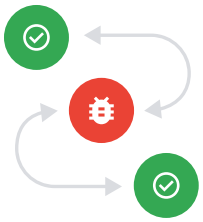
Profiter des protections des infrastructures sécurisées par défaut et du savoir-faire acquis en plus de deux décennies pour protéger les systèmes, les applications et les utilisateurs d'IA. Parallèlement, développer une expertise au niveau des organisations en vue de suivre le rythme des progrès de l'IA et de commencer à étendre et à adapter les protections des infrastructures en fonction de l'IA et de l'évolution des menaces. Par exemple, face aux techniques d'injection, telles que l'injection SQL, qui les menacent depuis quelque temps, les organisations peuvent adapter des mesures d'atténuation comme le nettoyage et la limitation des entrées pour mieux se défendre contre les attaques par injection de commandes.

2. Étendre la détection et la capacité de réponse afin d'intégrer l'IA aux stratégies de protection contre les menaces pesant sur les organisations

Détecter les cyberincidents liés à l'IA et y répondre rapidement grâce au renforcement de la surveillance des menaces et d'autres capacités. Pour les organisations, cela passe par un suivi des données entrant et sortant des systèmes d'IA générative visant à détecter les anomalies, et par le renseignement sur les menaces pour anticiper les attaques. Cela implique généralement une collaboration entre les équipes responsables de la confiance et de la sécurité, du renseignement sur les menaces et de la lutte contre les fraudes.

3. Automatiser les défenses pour s'adapter aux menaces nouvelles et existantes

Exploiter les dernières innovations d'IA pour renforcer et accélérer la réponse aux incidents de sécurité. Comme les adversaires utiliseront probablement l'IA pour augmenter leur impact, il est important d'en mobiliser les capacités actuelles et émergentes pour se protéger de façon flexible et rentable.



4. Harmoniser les contrôles au niveau de la plate-forme afin de permettre une sécurité homogène dans l'organisation

Mettre à jour les cadres de contrôle pour adopter l'atténuation des risques par l'IA, et étendre les protections à plusieurs plates-formes et outils afin que toutes les applications d'IA disposent des meilleures protections de façon flexible et rentable. Pour Google, cela consiste notamment à étendre les protections de sécurité par défaut aux plates-formes d'IA telles que Vertex AI et Security AI Workbench et à intégrer les contrôles et les protections dans le cycle de développement des logiciels. Grâce à des capacités destinées à répondre à des cas d'utilisation génériques, telles que l'API Perspective, l'ensemble d'une organisation peut bénéficier d'une protection de pointe.

5. Adapter les commandes pour permettre d'ajuster les mesures d'atténuation et instaurer des boucles de rétroaction plus rapides pour le déploiement de l'IA

Tester constamment l'application par l'apprentissage continu et adapter vos mesures de détection et de protection en fonction de l'évolution des menaces. Cela inclut l'utilisation de techniques telles que l'apprentissage par renforcement fondé sur les incidents et les retours d'utilisateurs. De même, cela inclut des mesures comme mettre à jour des ensembles de données d'entraînement, affiner les modèles pour une réponse stratégique aux attaques et permettre au logiciel utilisé pour la construction des modèles de contextualiser davantage la sécurité (par exemple, en détectant des comportements anormaux). Les organisations peuvent également réaliser régulièrement les exercices de simulation d'attaque de la Red Team afin de renforcer les garanties de sécurité des produits et des capacités fondés sur l'IA.

6. Contextualiser les risques liés aux systèmes d'IA dans les processus métiers adjacents

Réaliser des évaluations de bout en bout des risques portant sur le mode de déploiement de l'IA par les organisations. Ces processus chercheront notamment à évaluer le risque d'exploitation de bout en bout, notamment la traçabilité des données, la validation et le suivi du comportement de certains types d'application. En outre, les organisations devraient créer des vérifications automatisées visant à valider les performances de l'IA.