Google

Google のセキュア AI フレームワーク

AI の急速な進歩に伴い、効果的なリスク管理戦略も進化することが重要となっています。この進化を実現するために、Google では、AI システムを安全に保護するための概念的なフレームワークである、「セキュア AI フレームワーク（SAIF）」の導入を進めています。SAIF は 6 つの基本要素で構成されています。

1. 強固なセキュリティ基盤を AI エコシステムまで拡大する

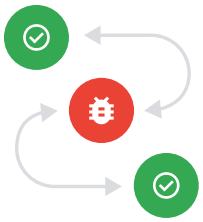
過去 20 年間にわたって構築されてきた、デフォルトで安全なインフラストラクチャ保護と専門知識を活用し、AI システム、アプリケーション、ユーザーを保護します。同時に、AI の進歩に対応するために組織内の専門性を高め、AI や脅威モデルに基づいたインフラストラクチャ保護の導入と強化を開始します。たとえば、SQL インジェクションのような攻撃手法は以前から存在しており、組織は入力値のサニタイジングや制限などの対策を適用することで、プロンプト インジェクション攻撃に対する防御を強化できます。

2. 検出機能と対応機能を拡張し、組織の脅威対策に AI を取り込む

脅威インテリジェンスやその他の機能を拡張することで、AI 関連のサイバー インシデントを適時に検知して対処します。組織にとって、これには AI 生成モデルの入出力を監視して、不正なデータや異常な動作を検出することや、脅威インテリジェンスを活用して攻撃を予測することが含まれます。通常、この取り組みには、セキュリティ、脅威インテリジェンス、不正行為防止チームとの連携が必要となります。

3. 防御を自動化し、既存および新規の脅威に対応する

最新の AI イノベーションを活用して、セキュリティ インシデントへの対応の規模とスピードを向上させます。攻撃者は AI を利用してその影響力を拡大させる可能性が高いため、組織にとっては AI に加えて、従来および最新の機能を活用して、敵対勢力に対する防御を迅速かつコスト効率よく行うことが重要です。



4. プラットフォーム レベルの管理を調整し、組織全体で一貫したセキュリティを確保する

AI リスクの軽減をサポートするコントロール フレームワークとの連携のほか、異なるプラットフォームやツールへの保護機能を拡張することにより、スケーラブルでコスト効率にすぐれた方法で、すべての AI アプリケーションで最適な保護を提供します。Google では、Vertex AI や Security AI Workbench のような AI プラットフォームにデフォルトで安全な保護を拡張することや、ソフトウェア開発ライフサイクルに制御と保護を組み込むことが含まれます。Perspective API など一般的なユースケースに対応する機能は、組織全体が最先端の保護を受けるのに役立ちます。

5. 管理を適応させて緩和策を調整し、AI デプロイ用に高速なフィードバック ループを作成する

継続的な学習を通じて実装を常にテストし、変化し続ける脅威環境に対処するために検出と保護を進化させます。その一例となる技術が、インシデントやユーザーからのフィードバックに基づく強化学習です。これには、トレーニング データセットの更新、攻撃に戦略的に対応するためのモデルの微調整、モデルの構築に使用されるソフトウェアがコンテキスト内にさらなるセキュリティを組み込む手順などが含まれます（異常な動作の検出など）。AI を活用する製品やその性能の安全確保を向上させるには、組織内でレッドチーム演習を実施する方法があります。

6. 周囲のビジネス プロセスにおける AI システムのリスクをコンテキスト化する

組織での AI 導入に関連するビジネス全体のリスク評価を実施します。これには、特定アプリケーションにおけるデータリネージ、検証、運用動作のモニタリングなど、エンドツーエンドでのビジネスのリスク評価が含まれます。さらに、AI のパフォーマンスの有効性を自動的に確認できる機能の構築も強く推奨しています。

Safer with Google