Google

Google 安全 AI 架構

AI 正在快速發展，重要的是有效的風險管理策略也要與時俱進。為了實現這個目標，
我們推出了安全 AI 架構 (SAIF)，這是一個安全 AI 系統的概念架構。SAIF 有六大核心要素：

1. 將強大的安全基礎延伸至 AI 生態系統

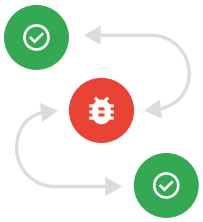
利用過去二十年建立的安全預設基礎架構防護技術和專業知識，保護 AI 系統、應用程式和使用者。同時，培養組織專業知識跟上 AI 技術的發展，開始在 AI 和瞬息萬變的威脅模型下增強並調整基礎架構防護。例如，像 SQL 插入攻擊這樣的插入技術已存在一段時間，而組織可適應變遷，比如輸入內容檢查和限制，以更妥善地防範提示詞插入攻擊。

2. 擴展偵測與回應，將 AI 引進組織的威脅範圍

透過強化威脅情報和其他能力，及時偵測並應對與 AI 相關的網路事件。對於組織而言，這包括監控生成式 AI 系統的輸入和輸出，以偵測異常情況，並利用威脅情報採取行動來防範攻擊。這項要務通常需要與信任與安全、威脅情報和反濫用團隊攜手合作。透過強化威脅情報和其他能力，及時偵測並應對與 AI 相關的網路事件。對於組織而言，這包括監控生成式 AI 系統的輸入和輸出，以偵測異常情況，並利用威脅情報採取行動來防範攻擊。這項要務通常需要與信任與安全、威脅情報和反濫用團隊攜手合作。

3. 將防禦自動化，跟上現有和新威脅的發展

善用最新的 AI 創新來提高應對資安事件的規模和速度。有心人士可能會利用 AI 來擴大其影響力，因此借助 AI 及其當前和新興能力來保持靈活度，並以符合成本效益的方式保護自己是非常重要的。



4. 協調平台層級控管機制，確保組織上下都有一致的安全性

協調控管架構來支援 AI 風險緩解，並擴充不同平台之間的防護措施和工具，確保所有 AI 應用程式都能以可擴充且符合成本效益的方式獲得最佳防護。在 Google，這包括將預設安全防護擴展到 Vertex AI 和 Security AI Workbench 等 AI 平台，以及在軟體開發生命週期中建立控管和防護措施。解決一般使用案例的功能 (例如 Perspective API) 可以協助整個組織從最先進的防護措施中受益。

5. 調節控管措施來調整緩解措施，並為 AI 部署建立更快的回饋循環

透過持續學習來不斷測試實施情況，並改善偵測和防護措施，以應對瞬息萬變的威脅環境。這包括基於事件和使用者意見回饋的強化學習技術，也涵蓋了不同步驟，如更新訓練資料集、微調模型以戰略性應對攻擊，並允許用於構建模型的軟體在情境中嵌入更多安全措施 (例如偵測異常行為)。各組織還可以定期進行紅隊演習，以提高 AI 產品和功能的安全保障。

6. 將 AI 系統風險置於業務環境流程中以便情境化

進行端對端風險評估，釐清組織部署 AI 時的相關風險。這包括評估端對端的業務風險，例如資料譜系、驗證和對某些應用程式操作行為的監控。此外，組織應該建立自動化檢查來驗證 AI 效能。

Safer with Google