Google

負責任的機器學習指南

指南

根據機器學習模型的輸出修改控制流程時請務必小心。

消耗機器學習原語輸出的工作流程可以修改為產生難以保護的攻擊面。與目前業界的建議一樣，請勿將資料納入執行工作流程中，除非是明確預期的工作流程。

避免直接存取模型和模型中繼資料。

與機器學習原語的直接互動可能會帶來許多不同的攻擊面；竊取、反轉等。讓這些物件遠離使用者，並允許受控的、僅經過核准的互動，並確保輸出附加到已建立的功能 (例如資料庫)。不要直接向使用者提供輸出內容。

確保模型在訓練和實施後採用完整性檢查

軟體的修改可以在建立和部署管道的多個階段進行，機器學習也是如此。檢查完後，再檢查一次。只執行您預期的內容：在操作之前進行認證。

功能的複雜性可能會讓控管越趨複雜。

保持互動簡短、切題、簡要。大型、複雜的工作流程在成功攻擊面代表更大的暴露風險。

從用於訓練的外延邏輯著手，然後在需要時轉向用於特徵和推論的內涵邏輯。

使用經過審查、已知良好的訓練集，是避免資料中毒和其他攻擊的好方法。然而，為了創造具有更強推論能力的特徵，請務必擷取新資料；根據已知良好輸入的預期結果，仔細測試這種訓練的結果，可以產生更具韌性的系統。

確保整合測試是根據原語的預期使用案例所建立。

與其他軟體一樣，機器學習工作流程需要整合和單元測試。機器學習軟體的啟發式本質可能會使建立可靠斷言變得困難，因此建議專注於工作流程期望的長期結果。

尋找使用機器合成的攻擊，特別是關注「你是什麼」的控制項。

在對聲音、影片和文字的真實性做出假設之前，務必問自己「現在機器能做到這個嗎？」，特別是當其用於提供存取權限時。

利用雙向政策維持經濟模型的領先地位，並確保強大、低成本的安全模型。

如果攻擊的成本低於後續補救的成本，則控管措施可能無效。 將安全成本制度化，根據成功攻擊的影響自動調整，逐步增加對攻擊者的成本。

Google

避免將在專屬資料上訓練的模型放在使用者控制的位置，例如裝置上。

不妨建立機器學習原語所涉及的使用案例，並假設理論上可以復原建立它的訓練資料。開發產品時，將其用於營運持續計畫和風險評估。

使用頻率限制來避免外推攻擊並防止不必要的資源消耗。

傳統控管措施仍然非常有效；在設計和導入機器學習技術時的層級控制，如存取控制清單、頻率限制和身分驗證。