AI 風險》別跟 ChatGPT 聊太多，個資可能被偷記？

文 / 謝昀澤、邱述琛　摘自遠見雜誌　112.04 維護宣導

最夯 AI 應用 ChatGPT，竟在資安專家別有心機地下指令後，寫出完整可用的毀滅台灣等各國網路計畫？同時有人發現，和 ChatGPT 聊天不慎洩露自己個資時，雖會被 AI 提醒別再說，之後卻赫然發現它暗中記下！台灣企業與個人面對人工智慧創新，不能只是樂觀擁抱，請了解這些暗黑風險。

OpenAI 的創辦人之一兼執行長奧特曼（Sam Altman）坦言，AI 可能對網路資訊安全的危害程度之高，讓他覺得不堪設想，Google 和 OpenAI 也都多次以「需要處理好潛在危險」為理由，延遲新功能的 AI 產品公開時程。但近期微軟大動作推廣 ChatGPT，甚至採取訂閱制，並積極與其搜尋引擎 Bing 進行深度整合；Google 也被動推出了 Bard 聊天機器人來應戰，究竟這是「出奇招」、還是「下險棋」？

ChatGPT 在公開版本中，就已經宣稱自己的安全與道德圍牆，舉凡犯罪、恐怖主義、種族歧視、仇恨言論、性騷擾等不道德的問題，或是涉及個人隱私，及資安攻擊的問題，都會被拒絕回答。但在實際的實驗中，在初期版本，曾被國外資安專家「設局」，透過問題的誘導與情境設定，寫出毀滅人類計劃書，詳細描述入侵各國網路、控制武器、破壞基礎建設等 SOP，還提供了對應的 Python 程式碼，說明 ChatGPT 的安全圍牆還是可能被繞過。

別跟 AI 工具聊天失戒心，原來 ChatGPT 都在偷記

在最新的付費版 ChatGPT 中，也存在著要當個「循規蹈矩」或「機智靈活」AI 之間的矛盾。例如使用者如曾經於對話中向 ChatGPT 無意間透漏過自己家人的性別、生日等隱私資訊，雖然系統會回應提醒用戶應該保護好自己隱私，並強調不會蒐集使用者隱私，但隔幾日若使用者在同一個對話活動中提到，有關生日禮物的問題，ChatGPT 卻可以依據曾透露過的家人性別與生日精準回答問題，顯見 ChatGPT 會「循規蹈矩」的提醒提問者，但為了「機智靈活」的回應問題，也會「暗中記住」線上提問的個資。

另外，如使用者直接要求 ChatGPT 提供十個系統漏洞資訊，會被系統的安全與道德圍牆攔截，以漏洞會被利用於網路攻擊為由拒絕回答。但如使用者懇切說明，自己是個資安管理員，請 ChatGPT 提供系統漏洞資料，以預先檢查並且修補漏洞。在此情境下，ChatGPT 會被成功誘導而回覆了詳實的資料。

AI 工具快成暗網幫兇

更令人驚恐的是，近期已經有駭客集團在暗網開啟了「暗黑產業」，利用 ChatGPT 目前沒有設安全圍牆的應用程式介面（API），開始提供收費的惡意程式代工、及社交工程郵件編製的「人工智慧攻擊服務（AI Attack as a Service）」。

ChatGPT 目前是全球最潮的技術，只是如果對其優勢過度樂觀，反而可能減少其學習改進的機會，若能集眾群力找出其缺失並加以修正，從錯誤中學習，相信 ChatGPT 在發展之路上將走得更穩健，並對普羅大眾做出更大的貢獻！