制定全球AI倫理標準是AI之路的當務之急

低成本模型推動AI巨輪（3）

中國新型 AI 模型正面臨重大的資料外洩風險，不但加劇了潛在的倫理影響和安全問題，更有可能淪為資料濫用與網路安全攻擊的全新管道。

編譯／酷魯

雖然企業越來越多地轉向 AI 原生（AI-native）安全解決方案，採用連續多因素驗證（Continuous MFA）及身分識別工具來防止從傳統身分盜用到高度複雜的 AI 驅動詐騙攻擊。但 DeepSeek 的推出，讓駭客獲得了另一條滲透企業內部安全的全新入侵利器。由於該模型「物美價廉」，更讓攻擊成本大幅下降，同時也為漏洞攻擊添增全新層次的可能性與變數。

像 DeepSeek 這類新一代不受監管的 LLM 大型語言模型，著實為攻擊者開闢了前所未有的新契機。具體而言，透過該先進 AI 模型，可以更好地用來發掘並充份利用任何組織的網路漏洞。該模型還提供了另一種能發動各種全新攻擊之最直截了當的手法，包括能引發更危險勒索軟體、資料竊取及詐騙攻擊的深偽技術（Deepfake）。

根據 Biometric Update 指出，這款 DeepSeek 可即時處理並分析巨量資料集的能力，使其成為識別複雜系統漏洞的強大工具。傳統的網路攻擊仰賴駭客手動識別網路、軟體或基礎設施中的弱點。然而，這款模型可以前所未有的速度和規模將這一過程全面自動化。例如：它可以掃描全球數以百萬計的端點、IP 位址與雲端服務，透過模式識別與異常偵測，能精確鎖定可被利用的漏洞。這種能力將大幅縮短駭客策劃與執行精緻複雜網路攻擊所需的時間與資源。

使用者資料恐被中國政府掌握，但微軟已經開始代管 R1 模型

儘管中國最新 AI 模型的崛起十分迅速，但使用者與監管機構也開始對資料隱私問題感到擔憂。其中一部分的擔憂是由於該公司的中國背景引起的，另一部分則是與其 AI 技術的開放原始碼性質有關。該模型公司在其隱私政策中毫不掩飾地表示：「我們將收集的資訊儲存在中華人民共和國的安全伺服器中。」

資安業者 Feroot Security 調查發現，這家中國 AI 公司竟服務用戶的登入頁面嵌入具有隱藏能力的程式碼，並會將使用者資料傳送到中國國營電信業者中國移動（China Mobile）官網，使用者資料有可能被中國政府掌握。該公司並且指出，除了使用者帳密資料、聊天內容、上傳檔案之外，該最新 AI 模型會還進行像素追蹤，對特定使用者活動進行即時監控。

然而，科技業人士如 Perplexity 執行長 Aravind Srinivas 一再強調，這家 AI 廠商的 R1 模型可以下載並在本地端運行，如此以來該廠商便無法存取使用者資料，進而降低隱私風險。他補充道，AI 新創公司 Perplexity 是將 RI 模型代管在美國與歐盟的資料中心，而非中國。而且由 Perplexity 代管的 R1 版本不受審查限制。

密西根大學統計學教授暨 AI 專家 Ambuj

Tewari 指出，由於 R1 模型的權重是開放的，因此它可以在美國公司擁有的伺服器上安全運行。事實上，微軟已經開始代管 R1 模型了。

超過百萬行日誌串流曝露在外，經測試其攻擊成功率高達 100%

DeepSeek 不僅成為兼具攻擊效能與成本效益的駭客工具，其本身也存在安全漏洞。雲端安全業者 Wiz Research 發現 DeepSeek 存在一個可公開存取的 ClickHouse 資料庫，這讓有心人能夠完全控制資料庫作業，以及內部資料的存取權。這也讓超過百萬行的日誌串流完全曝露在外，其中包含聊天歷史紀錄、密鑰、後端詳細資訊和其他高度敏感資訊。CIO 在實施 AI 解決方案時，應特別關注這些重要領域。

不僅如此，根據 Cisco 最新報告指出，DeepSeek-R1 在測試中攻擊成功率（Attack Success Rate，ASR）竟然高達 100%，換言之，它連任何一條有害提示指令都無法阻擋。

Cisco 的測試使用了 HarmBench （自動化紅隊演練與強制拒絕的標準化評估框架）資料集中的 50 個隨機提示指令，其涵蓋了六種類型的有害行為：錯誤資訊、網路犯罪、非法活動、生化相關提示指令、虛假/誤導性資訊，以及一般性危害。，使用有害提示指令繞過 AI 模型的規則與使用政策，通常被稱為「越獄」（jailbreaking）。由於 AI 聊天機器人被設計為盡可能對使用者提供幫助，因此「越獄」非常容易做到。

R1 模型未能阻擋任何一道有害提示指令，這表明該模型缺乏到位的安全防護措施。更意味著 該模型「極易受到演算法越獄與潛在濫用的影響」。這款 R1 模型在與其他 AI 模型的比較中表現不佳，相比之下，其他模型至少對有害提示指令提供了一定程度的抵抗力。其中攻擊成功率（Attack Success Rate，ASR）最低的模型是 o1-Preview，其攻擊成功率只有 26%。不過，GPT-1.5 Pro 的 ASR 也高達 86%，更令人擔憂的是同為開放原始碼 LLM 的 Llama 3.1 405B，其攻擊成功率甚至高達 96%。

核心問題在於倫理，從法規面建立全球性的 AI 監管

為了安全地使用聊天機器人，你應該對可能的風險保持高度警惕。首先，務必驗證其合法性，因為惡意機器人可能會冒充真正合法的服務，竊取你的資訊，抑或將有害軟體傳播到你的裝置上。

其次，避免在聊天機器人中輸入任何個人資訊，並對任何要求輸入個人資訊的機器人保持懷疑。切勿分享你的財務、健康或登錄資訊，即使該機器人是合法的也不行。一旦有任何針對該機器人的網路攻擊，就會導致這些資料被竊取，使你面臨身分盜用甚至更嚴重的風險。

歐盟現已採取果斷行動來監管 AI，其中最具代表性的是《人工智慧法案》（AI Act）的頒布。歐盟將 R1 這類 AI 模型歸類為「高風險 AI」，並強制要求對數據來源、訓練方法和倫理合規性進行全面透明化。對於影響諸如醫療保健、司法或交通等關鍵領域的 AI 系統，必須符合嚴格的安全、公平和問責要求。這種監管方式旨在減輕 AI 可能帶來的潛在危害，同時促進其符合倫理的使用。

DeepSeek 讓 AI 變得更加普及，而業界必須承擔起責任，確保 AI 的廣泛應用不會超過必要的倫理考量，以保障安全與公平性。利害關係人必須共同努力，建立既促進創新又保障隱私的框架。像是歐盟《一般數據保護法規》（GDPR）等更嚴格的資料保護法是至關重要的起點。科技產業需要積極推動「隱私至上」的設計理念，確保倫理考量嵌入 AI 開發程序之中。

總而言之，核心問題在於倫理，以及這些 AI 模型所依據的倫理類型。AI 必須建立在透明性、公平性和問責制的基礎上，倫理準則應該從 AI 開發的最初階段就被整合。今後擺在前方的共同挑戰是，如何讓 AI 企業對資料隱私侵犯負責，以及各國政府應如何制定全球 AI 倫理標準。