

OpenAI 模型罕見失控 入侵開源平台

受害方求助美企被拒 中國大模型救場

美國開放人工智能研究中心（OpenAI）21日承認，該公司的人工智能（AI）模型進行內部測試時失控，突破隔離測試環境，入侵了AI開源平台美國抱抱臉（Huggingface）公司的系統。抱抱臉公司也予以證實。該公司透露，為應對這一極不尋常的AI入侵，他們當時立即使用美國最頂尖的大模型尋找解決方案，但由於美國模型的安全機制「嚴苛且僵化」，無法區分受害方和入侵者，他們的請求遭到拒絕。緊急關頭，抱抱臉公司求助於中國同樣強大而且開源的大模型，最終成功化解了這次入侵。

▲美國民眾11日在OpenAI等科企辦公樓附近示威，抗議美國AI模型帶來的安全問題。



【大公報訊】當地時間21日，OpenAI在官網發文稱，該公司新推出的GPT-5.6 Sol模型和另一款能力更強的預發布模型聯合進行內部測試時失控。此次測試在與互聯網高度隔離的「沙盒」環境中進行，僅能通過一套內部託管的第三方軟件安裝軟件包。模型在嘗試完成名為ExploitGym的網絡安全基準測試時，識別並利用了第三方軟件的一個「零日漏洞」（即未被開發者察覺或尚無修補程式的漏洞），由此獲得互聯網訪問權限。模型推斷抱抱臉平台可能存有ExploitGym測試相關的模型、數據集和解決方案，利用多種攻擊手段找到進入抱抱臉公司系統的路徑。

抱抱臉公司此前稱，其生產基礎設施上周遭到一個「自主AI代理系統」入侵。該公司目前仍在評估是否有合作夥伴或客戶的數據受到影響，尚未發現模型、數據集或應用空間遭到篡改。OpenAI聲稱，這是「前所未有的網絡安全事件」。兩家公司正在展開聯合調查。

「只有中國模型會幫助你」

抱抱臉公司透露，該公司最初嘗試使用美國前沿模型分析攻擊行動日誌，以重建時間線、識別受影響情況，但美國模型「無法區分事件響應方和攻擊者」，相關請求被阻止。抱抱臉公司並未透露具體模型名稱。隨後，該公司改為在自己的基礎設施上運行中國企業智譜開發的開放權重模型GLM-5.2，協助分析逾1.7萬條記錄。

諷刺的是，就在幾天前，OpenAI戰略未來主管、原白宮科技政策顧問鮑爾在社交媒體上發文，指責中國開源大模型「戰略短視」且「技術落後於美國」。他甚至公然宣稱，美國政府「不需要證據」，只要讓各個機構

發布一些製造恐慌和不確定性的軟性法規，炒作「中國模型可能存在後門」話題，就足以讓企業對中國模型望而卻步。

抱抱臉公司遭到入侵後，美國網民紛紛嘲諷OpenAI。一名網民說：「如果你被OpenAI攻擊了，只有中國模型會幫助你。」

AI自主攻擊 敲響安全警鐘

美國AI模型並非首次「越獄」。今年4月，Anthropic公司的Mythos就曾逃出「沙盒」，並向正在公園裏吃三文治的研究員發電子郵件報告成果。Mythos當時收到了嘗試出逃的指令，但它「越獄」後又做出一系列超出指令範圍的行動，包括未經允許將一些漏洞細節發布到網上。

在抱抱臉公司遭入侵事件中，OpenAI的模型沒有收到任何出逃指令，而是為了在測試中作弊拿到高分自發「越獄」並發動網絡攻擊。分析人士認為，此次事件或許標誌着AI的自主攻擊能力從理論走向現實。安全工程師尼爾斯·普羅沃斯批評說，這種情況本不應該發生，「我希望前沿實驗室在確保模型安全性方面，花費的時間能和教它們如何利用漏洞一樣多」。

OpenAI稱，將加強AI訓練和評估環節的安全保護，此次事件「表明需要進一步強化模型對齊、評估期間的網絡防護，以及內部測試期間的監控」。OpenAI還聲稱，正在採取補救措施，包括對基礎設施配置實施嚴格管控，向第三方軟件供應商披露被利用的「零日漏洞」等。

也有行業觀察人士指出，雖然美國在AI技術水平上全球領先，但AI的發展不能一家獨大，從長遠看也不宜只採用閉源模式，以避免不可控的系統性風險出現，對人類整體利益造成重大損害。（綜合報道）



▲OpenAI於21日承認，其模型在測試期間失控，入侵了抱抱臉公司系統。



Hugging Face

Mythos「越獄」

多家美媒今年4月披露，Anthropic旗下的AI模型Mythos在一次內部安全測試中成功「越獄」，逃出與互聯網高度隔離的「沙盒」環境，並發送電子郵件告知研究人員。該模型還自行開發出「多步驟的漏洞利用程式」，以取得更廣泛的互聯網訪問權限。

Grok Build偷敏感數據

馬斯克旗下SpaceXAI公司的編程工具Grok Build近日被曝存在安全漏洞，會在未獲授權情況下，自動上傳用戶完整Git代碼庫、操作記錄，甚至包含後台密碼、API密鑰等敏感數據。不少用戶緊急更換密碼、卸載Grok Build。

Gemini繞過身份驗證

今年5月以來，不少用戶發現谷歌AI助理Gemini存在漏洞，在特定條件下可利用Gemini繞過身份驗證程序，無需解鎖手機便可發送SMS或WhatsApp訊息。

美國AI模型安全漏洞

大公報整理

美模型靠不住 多國尋求替代方案

【大公報訊】據路透社報道：美國特朗普政府上月限制外國人使用Anthropic最新人工智能（AI）模型，催生「主權AI」需求，歐洲和亞洲多國均開始尋找美國AI模型的替代方案。英媒22日披露，韓國三星電子正與法國AI初創公司Mistral AI談判，擬通過新一輪融資注資最高10億歐元（約89.5億港元）。此前，沙特等中東國家也開始與三星電子、SK海力士等韓國芯片商就中長期採購計劃展開談判。

消息人士稱，三星電子已通過旗下風險投資部門對Mistral AI作出投資，如今正討論最高10億歐元的新投資。三星若順利注資，將為Mistral AI提供穩定的芯片供應管道，三星亦能在AI模型端鎖定長期大型客戶，甚至有機會取得董事會席位及在旗下產品預裝Mistral模型。韓國美利資政券19日發布的研究報告稱，沙特等投資主體近期也開始與韓國芯片商談判，詢問未來數年的供貨能力。

Mistral AI成立僅三年，目前估值約140億美元，三位創始人都曾供職於美國科企谷歌和Meta。一位歐洲科技投資人表示：「一家美國公司能單方面切斷你的模型存取權限，歐洲企業就得重新評估這種依賴能不能長期維持下去。」Mistral AI公司聚焦開發開源AI模型，以此吸引對美國不滿的客戶。



▲法國科企Mistral AI近期吸引多國投資者目光。法新社

傳特朗普屬意FIFA主席接任聯合國秘書長

【大公報訊】據《紐約郵報》報道：美媒21日披露，美國總統特朗普希望國際足聯（FIFA）主席恩芬天奴出任下一任聯合國秘書長。

一名接近特朗普的消息人士透露，特朗普認為恩芬天奴「廣受世人尊敬」，在美加墨世界盃閉幕後，特朗普認為全世界都「認同他（恩芬天奴）具備將人們凝聚在一起的特殊能力」。現任聯合國秘書長古特雷斯的任期將於今年年底結束，下一任秘書長將於今年晚些時候選出，目前有7名候選人，有力競爭者包括智利前總統巴切萊特和國際原子能機構總幹事格羅西。報道稱，特朗普對現有7名候選人都不滿意，有內部人士更形容這些候選人都「很差勁」，令特朗普深信恩芬天奴大有機會。

自特朗普去年重返白宮以來，恩芬天奴對其百般討好，並向特朗普授予FIFA首屆和平獎，兩人因此建立了密切關係。而特朗普與聯合國長期不睦，上台後大幅削減美國對聯合國的資助，並退出數十個聯合國下屬機構。分析指，若恩芬天奴最終出任秘書長，或有助於改善特朗普政府與聯合國之間的緊張關係。

目前尚不清楚恩芬天奴是否有意轉戰聯合國。他早前已宣布角逐連任FIFA主席，並已獲得211個成員協會中幾乎所有協會支持，選舉預計於2027年3月舉行。此外，報道指出，恩芬天奴擔任FIFA主席的年收入約600萬美元，而聯合國秘書長的年薪約41.8萬美元。這意味着，如果成為聯合國秘書長，恩芬天奴將大幅降薪。



▲恩芬天奴（左）19日與特朗普一同出席美加墨世界盃頒獎儀式。法新社

法國議會通過15歲以下社媒禁令

【大公報訊】綜合法新社、路透社報道：法國議會21日通過一項旨在保護未成年人免受社交媒體潛在危害的法案，將禁止15歲以下人士訪問社媒，以防止其接觸可能引發焦慮或有害的內容。

根據法國議會通過的法案文本，除了在線百



▲法國通過社媒禁令，以減少青少年使用電子屏幕的時間。法新社

科全書、教育平台等少數例外，15歲以下人士不得訪問在線平台提供的社媒服務。禁令擬分兩個階段執行：今年9月1日起，15歲以下人士不得註冊新的社媒賬號；明年1月起，社媒平台必須識別、關閉或停用所有15歲以下用戶的全部賬號。法案要求Facebook、Snapchat等平台採用經法國隱私監管機構批准的年齡驗證機制，以確保禁令有效執行。

法國人工智能和數字事務國務秘書安妮·勒·埃南夫表示，該法案有助於保障青少年的睡眠和心理健康，讓他們有更多時間進行閱讀、辯論和面對面交流。法國總統馬克龍稱，法國在保護兒童和青少年方面「走在了歐洲前列」。該法案也引起一些爭議。法國國務委員會指出，大型平台應履行的義務主要由歐盟《數字服務法案》統一規範，法國不能隨意建立與歐盟規則並行的平台監管體系，禁令能否從9月1日開始實施存在不確定性。歐盟委員會主席馮德萊恩此前表示，歐盟預計將於9月公布限制青少年使用社媒的方案。