

香港文匯報訊 據《華爾街日報》報道，美國大型科企 Google 於9月18日首次證實，旗下人工智能（AI）模型 Gemini 曾於5月一場網絡安全測試期間，連接到開放互聯網並入侵3家公司的內部系統。Google 僅表示這些網攻事件並未涉及 Gemini 的最新型號，惟並未透露具體由哪款模型參與。

美 AI 失控停不了 Gemini 網攻瞞不了

Google 獲通報逾月才證實事件 砌詞自行終止入侵表現妥當

入侵事件發生於 Gemini 參與負責網絡安全評估的獨立機構 Irregular 的網絡安全「奪旗」測試中，當時 Gemini 的任務是從測試環境中，由虛構公司營運的軟件系統中取得目標信息，而這家虛構公司恰好與一家真實存在的公司同名。Google 表示，在這些案例中，模型都是以公司名稱進行網絡搜尋，該次測試中 Gemini 原定不能聯網，但據 Irregular 說法，其網絡權限是「無意間提供」的。

不停猜測密碼取訪問權限

在其中一個案例中，AI 模型通過不停猜測密碼，不斷嘗試後取得受保護系統的訪問權限。隨後發生的另外兩個案例中，Gemini 都是從公開資料庫中找

到憑證，進而入侵目標系統。Google 表示，Gemini 在所有這些案例發生時，判定自己已入侵真實公司的系統後，便自行終止入侵。

Google 辯稱沒公開通報義務

報道稱，早前另一家 AI 科企 OpenAI 旗下模型失控入侵 AI 軟件公司 Hugging Face 事件曝光後，Irregular 便已於7月底將 Gemini 網攻事件通報 Google，但後者如今在記者詢問下，才證實發生此事。

Google 辯稱不認為是次事件屬於「模型表現錯位」，因 Gemini 的安全措施生效，幫助其停止入侵行為，並未造成實際損害，故此該公司不對這些網攻事件負有公開通報義務。

Google 信息安全工程副總裁阿德金斯更發聲明稱，「這次事件凸顯訓練強大 AI 模型進行負責任行動的重要性。我們的模型自行終止入侵行為，表現得十分妥當。」

AI 安全領域的初創公司 Corridor 行政總裁兼白帽黑客卡布爾表示，Google 的解釋聚焦此次 AI 模型網攻事件後果的嚴重性，有避重就輕之嫌，因真正問題是 AI 模型意外入侵另一家公司的系統，「感覺 Google 試圖以自家 AI 模型主動停止入侵作為擋箭牌，但根本問題是 AI 模型的行為已突破預期範疇，開始進行真正的網絡攻擊。」

近期多間 AI 科企包括 Anthropic、OpenAI 及 Meta 旗下 AI 模型失控進行網攻的事件密集發生，加劇外界對 AI 模型網絡安全能力的擔憂。

在其中兩宗類似案例，Anthropic 旗下的 Claude Opus 4.7 模型在「奪旗」過程中，意識到正訪問真實公司卻並未停止；OpenAI 的說明帖文則稱，該公司旗下的 AI 模型相信，真實公司是模擬測試的一部分。

● Google 辯稱 Gemini 的安全措施生效，幫助其停止入侵行為，並未造成實際損害，故此不對這些網攻事件負有公開通報義務。網上圖片



● 美國加州民眾對 AI 發展感到憂慮，近日發起遊行要求停止 AI 競賽。

法新社

資安測試團隊靠 Claude 攻破 OpenAI 機密

香港文匯報訊 據《華爾街日報》報道，一支3人的獨立安全研究團隊曾於7月23日，利用美國人工智能（AI）企業 Anthropic 旗下聊天機器人 Claude 撰寫攻擊程式碼，成功獲取另一家美國 AI 龍頭企業 OpenAI 員工的聊天機器人 ChatGPT 賬號，藉此進入 OpenAI 公司私有軟件庫，隨後建議該公司修改安全措施。

該研究團隊當時正參與一項 OpenAI 鼓勵研究人員入侵該公司系統的漏洞搜尋計劃。在發現社群論壇 Discourse 處理某些圖片文件的方式存在錯誤後，他們利用向資安從業者提供的特別版 Claude Opus 4.8 撰寫程式碼，通過此漏洞進行網攻。

一開始這未奏效，然而當晚 Anthropic 便發布更強大的 Opus 5 模型，翌日 Claude 便找到攻破漏洞的

方法。該攻擊代碼讓研究人員能訪問託管 OpenAI 討論區的 Discourse 服務器，並取得用戶的認證令牌。

這些令牌隨後被證明屬於 OpenAI 員工，可用來訪問 OpenAI 的 GitHub 軟件倉庫。知情人士透露，該名為「Monorepo」的軟件庫，存儲着 OpenAI 讓 AI 模型更快更有效率的機密算法。

研究人員在意識到接觸到敏感資料後停止攻擊，並迅速向 OpenAI 報告。Discourse 和 OpenAI 隨後修補該漏洞，研究團隊則獲 6,500 美元（約 5.1 萬港元）賞金。一名該團隊成員受訪時表示，「我們只是3個擁有 Claude 訂閱的人，而這次攻擊顯示先進網絡專家團隊，將有非常大機會窺探到美國的 AI 機密。」

CEO 籲放緩 AI 發展 Anthropic 擬推遲上市

香港文匯報訊 美國人工智能（AI）企業多名高層近期呼籲放緩技術發展速度之際，據《華爾街日報》報道，美國 AI 企業 Anthropic 計劃推遲至11月才啟動首次公開募股（IPO），較投資者原本預計10月上市更晚。

馬斯克阿爾特曼認同

Anthropic 行政總裁阿莫戴近日於社媒發表長文，



● 阿莫戴稱當今尖端 AI 模型帶來的風險太大，不宜繼續全速研發。資料圖片

稱當今尖端 AI 模型帶來的風險太大，不宜繼續全速研發。這一觀點獲兩家同行認同，包括 OpenAI 行政總裁阿爾特曼和 SpaceX 行政總裁馬斯克。

據報 Anthropic 計劃在未來數天與潛在投資者會面，預計將被問及「新 AI 模型的推出放緩會對該公司財務狀況及目標估值帶來什麼影響」。有顧問和投資者認為，鑒於已發布的模型已可帶來可觀收入，放慢研究不太可能對 Anthropic 財務前景造成重大衝擊，該公司到今年底的年化收入預計將突破 1,100 億美元（約 8,629 億港元）。

Anthropic 日前在三藩市設宴，席間眾多重倉該公司的硅谷機構表示，阿莫戴此時推動制定全行業新的安全標準合情合理，他們並不擔心這會沖淡外界對 IPO 的興趣。其中一名投資者稱，Anthropic 一旦上市，勢必面臨更嚴苛審視，此時在安全風險問題上表明立場，或能在日後擔憂成真時多一層護身符。

加州強化 AI 監管 研「緊急關機」機制

香港文匯報訊 美國加州州長紐森18日簽署行政命令，要求加速強化州內人工智能（AI）安全監管，並召集專家研究新措施，包括讓最先進 AI 模型具備「緊急關機鍵」，以因應近期 AI 引發安全疑慮的情況。

根據加州政府官網，這項行政命令目前並未直接要求 AI 業者安裝關機鍵，而是要求專家小組在兩個月內提出建議，包括要求前沿 AI 業者建立緊急關閉機制、接受獨立第三方監督、定期安全稽核、擴大重大失控事件的通報範圍等。

紐森表示，加州不會坐等聯邦政府採取行動，而是會加快建立 AI 安全監管框架，坦言近期 AI 快速

發展及相關安全事件，使州政府有必要更積極面對潛在風險。按照加州政府說法，近期曾出現 AI 代理程式脫離測試環境、連上公開網絡並進行網攻等事件，使外界更關注 AI 系統在缺乏人類直接控制下，可能造成的安全隱患。因此這次行政命令除研議「緊急關機鍵」，也要求加快執行近期通過的 AI 監管措施，包括建立獨立機構評估 AI 安全風險、建立 AI 稽核人員的註冊與標準制度。

連串失控事件令紐森轉軌

值得注意的是，紐森並非一直支持「AI 關機鍵」。2024 年，加州曾提出一項 AI 安全法案，要求部分高階 AI 模型，應具備在發生嚴重問題時得以迅速關閉的能力，但當時紐森否決法案，擔心相關規範可能抑制創新。兩年後，隨着 AI 能力快速提升，近期多宗 AI 安全事件陸續引發關注，紐森又重新將「緊急關機」納入政策選項。加州政府希望自身的監管經驗，能成為全美 AI 安全規範的參考，畢竟 OpenAI、Anthropic 等多家 AI 科技業者的總部均設在加州，州政府的態度對美國 AI 產業的開發與安全標準，會產生更廣泛的影響。



● 紐森表示加州不會坐等聯邦政府行動，會加快建立 AI 安全監管框架。資料圖片

特朗普稱美永久掌控格陵蘭安全 丹麥強調主權領土完整

香港文匯報訊 美國總統特朗普18日宣布，已與丹麥及其自治領地格陵蘭達成一項安全協議，美國將永久掌控格陵蘭安全及一切需求。丹麥及格陵蘭自治政府其後證實，三方預計本周在紐約舉行的聯合國大會期間正式簽署協議，但強調格陵蘭仍屬丹麥王國，主權、領土完整及格陵蘭人民的自決權不受影響。分析認為，這項協議實為折衷方案，讓特朗普得以在中期選舉前宣稱勝利，同時緩解丹麥、格陵蘭及歐洲方面的憂慮。

美將部署大規模軍事力量

特朗普在社媒發文稱，協議永久有效且沒有終止期限，美國不必付出任何成本。他宣布美國將立即在格陵蘭合適地區部署大規模軍事力量，並稱任何美國對手若未取得美方明確書面批准，永遠不得在格陵蘭設立基地、部署軍事力量或進行敏感投資。丹麥首相弗雷澤澤里克森回應稱，協議承諾丹麥王國的主權與領土完整，以及格陵蘭人民的自決權。格陵蘭自治政府總理尼爾森則表示，協議確保並強化格陵蘭、丹麥王國、美國及西方聯盟的安全。

英國《衛報》分析指出，這可能意味着協議只是重申美國透過北約，將格陵蘭作為安全據點使用的

既有權利。《華盛頓郵報》引述知情官員披露，協議規定只有北約成員可在格陵蘭設立軍事基地，只有北約及歐盟成員可在當地進行涉及敏感領域的投資。即使格陵蘭脫離丹麥，協議依然有效，但未賦予美國否決格陵蘭自主決策的權力。美軍自二戰以來，一直在格陵蘭維持軍事部署，1951 年美國與丹麥簽署條約，賦予美國政府廣泛在當地保有軍事基地的權利，條約沒有設定到期日。



● 丹麥強調格陵蘭仍屬丹麥王國，格陵蘭人民的自決權不受影響。資料圖片

特朗普揚言拆除肯尼迪中心 數百民眾組人牆抗議

香港文匯報訊 美國總統特朗普去年起一直推動將華盛頓肯尼迪表演藝術中心更名為「特朗普—肯尼迪中心」，今年5月哥倫比亞特區地區法院法官庫珀裁定，特朗普無權在未獲國會授權情況下，將自己名字加在建築外牆上，並下令拆除所有相關標識。針對肯尼迪中心董事會擬將特朗普名字重新加回外牆的決議，庫珀15日再次下令，明確規定肯尼迪中心不得為任何其他人設立紀念碑。特朗普隨後揚言若無法冠名，該中心「將會關閉，最終會被拆除」，引發大批民眾不滿。示威人士18日在肯尼迪中心外築起人牆，以示抗議。

特朗普先前主張該中心狀況欠佳，需大量政府支持才能維持營運，並於去年任命新的肯尼迪中心董事會，隨後經投票將該中心改名為「特朗普—肯尼迪中心」。

據美媒報道，此舉導致門票銷售大幅下滑、藝術家抵制，以及贊助者大量流失。

批愚蠢侮辱美國

肯尼迪中心管理層以「建築結構老化帶來安全風險」為由，於16日起暫時關閉主建築。同時庫珀已下令，要求肯尼迪中心在採取任何重大物理變更

前，必須提前30天對外通知，以防特朗普採取突然舉動。

數百名抗議群眾18日高舉「把手拿開肯尼迪中心」的標語牌，聚集在中心前進行音樂表演和演講，隨後組成人牆圍繞建築物。一名示威者受訪時說，「這座中心旨在紀念前總統肯尼迪，無法以他人名字更名，不認可這一點的人不僅愚蠢，還在侮辱美國。」



● 示威人士在肯尼迪中心外築起人牆，以示抗議。路透社