

美國人工智能頻繁失控 監管成疑

谷歌 AI 模型被爆入侵三家公司

當地時間 18 日，美國科技巨頭谷歌公司證實，旗下人工智能（AI）模型 Gemini 在今年 5 月的一次網絡安全能力測試中，入侵了三家外部真實公司的系統。據悉，谷歌早在 7 月份已知悉相關事件，但一直未對外公開。分析指出，這是繼 OpenAI、Anthropic 和 Meta 之後，美國再有知名 AI 模型失控，引發外界對 AI 安全和監管的擔憂。

【大公報訊】谷歌公司安全工程副總裁阿德金斯 18 日發表聲明說，上述入侵事件發生在今年 5 月進行的常規演練評估中，由以色列網絡安全評估公司「非常規」（Irregular）負責。在其中一項測試中，Gemini 的 AI 智能體通過反覆嘗試猜中密碼，成功進入一個真實世界公司的系統。在另外兩項測試中，AI 從公開存儲庫中找到登錄憑證，並利用這些憑證進入其他公司的系統。

「非常規」稱，相關測試場景使用虛構公司作為目標，但這家虛構公司與現實中的某家企業同名。同時，測試環境意外開放了互聯網訪問權限，導致 AI 模型將真實網絡目標誤認為測試場景的一部分，並採取網絡攻擊行動。

研究員用 Anthropic 模型攻破 OpenAI 系統

這是谷歌 AI 首宗失控入侵事件，由《華爾街日報》在 19 日率先曝光。據悉，谷歌 7 月底已獲悉情況，當時正值 OpenAI 智能體被爆入侵開源平台「抱抱臉」之際，但谷歌在媒體問詢前，並未主動對外公開事件。阿德金斯稱，谷歌已向被入侵的公司通報了相關事件，強調 Gemini 的 AI 智能體發現入侵真實機構後，就停止了相關入侵行動，沒有造成任何損失。AI 安全初創公司 Corridor 行政總裁凱布爾批評谷歌迴避根本問題，並強調「公眾有權知道這些」。

另外，《華爾街日報》18 日報道，美國網絡安全公司 Hacktron 的三名研究人員在今年 7 月利用 OpenAI 社區論壇的託管平台 Discourse 上的漏洞，透過 Anthropic 旗下的 Claude 模型編寫攻擊代碼，成功入侵了 OpenAI 內部系統，並取得 OpenAI 一名員工的 ChatGPT 及 Codex 賬戶控制權，藉此連接至 OpenAI 內部程序碼庫。整個攻擊流程由發現漏洞到攻入內部系統僅花 72 小時。事件屬合法漏洞懸賞測試，OpenAI 向 Hacktron AI 發放 6500 美元獎金。事件也顯示 AI 模型的黑客能力在大幅提升。

四大 AI 巨頭涉嫌壟斷被起訴

針對 AI 模型能力大幅提升造成的安全風險，Anthropic 行政總裁阿莫戴 12 日率先呼籲「管控前沿 AI 的發展」，獲得 OpenAI 行政總裁阿爾特曼、SpaceX 創辦人馬斯克和谷歌 DeepMind 創辦人哈薩比斯的公開支持。

Politico 報道，Anthropic、OpenAI、SpaceX 與谷歌 18 日在北加州的聯邦法院遭遇民事訴訟，四名原告指控這 4 家公司可能透過公開表態，通過商業協調放緩 AI 發展，涉嫌違反美國反壟斷法。原告代表律師表示，興訟目的在於避免全球大型科技企業私下達成協議，影響 AI 發展與相關安全措施。原告方認為，涉及人類持續風險的防護規範，應由政府以透明且合法的方式制定，同時接受公眾監督。專家也指出，由少數幾家企業自行界定何為「危險 AI」並自行挑選核查人員，「不是真正意義上的監管」。

據悉，美國政府內部對如何監管 AI 仍未形成共識。白宮內部圍繞 AI 監管形成兩派對立，白宮幕僚長威爾斯、財長貝森特及國家網絡總監凱恩克羅斯等人近數月幾乎每日開會，試圖說服特朗普加強政府介入。白宮 AI 事務顧問薩克斯、英偉達 CEO 黃仁勳以及 Meta 創始人朱克伯格等人則頻頻向特朗普進言，力主維持寬鬆監管，避免拖慢美國 AI 發展。

聯合國安理會 23 日將開會討論 AI 與國際安全問題，阿爾特曼將出席。會議召集方法國指出，考慮到 AI 被惡意濫用的風險，有必要迅速採取行動，評估失控和被濫用的風險。

（綜合報道）

特朗普延長 H-1B 簽證 10 萬美元收費一年

【大公報訊】綜合路透社、美聯社報道：美國白宮 18 日宣布，總統特朗普已簽署行政令，針對 H-1B 工作簽證收取 10 萬美元（約 78 萬港元）費用的新措施將延長一年至 2027 年 9 月 21 日。該措施原定本月底到期，此舉預料再次對美國企業產生重大影響。

根據去年 9 月特朗普簽署的行政令，僱主為每名透過 H-1B 計劃從海外聘用的技術工人提交申請時，須繳納 10 萬美元費用，除非獲得國土安全部的「國家利益豁免」。該費用不適用於已持學生簽證在美的外籍人士，也不適用於現有 H-1B 續簽申請。特朗普聲稱，延長該措施將「持續保護美國的經濟和國家安全利益」。特朗普還要求多個部門合作，加強對 H-1B 申請審查力度，包括須考慮僱主近期或計劃對美國工人實施的裁員情況。

上述高額收費政策目前面臨法律挑戰。今年 6 月，美國聯邦地區法院法官裁定該政策違法，認定 10 萬美元收費是「未經國會授權的稅收」。美

國政府已提出上訴。此外，8 月 24 日，美國國土安全部在《聯邦公報》上公布新規，提議對 H-1B 申請收取 103265 美元費用。新規有 30 天的公眾諮詢期，9 月 24 日到期。

受最新 H-1B 簽證政策影響，美國國土安全部在 2025 年 9 月至 2026 年 2 月，僅收到 85 份符合條件的 H-1B 申請並收取相關費用，申請數量比前一年減少了 87%。蘋果、微軟和谷歌等美國科企，目前已轉為在海外建立辦公室並直接在當地招募人才。



▲美國總統特朗普 18 日在白宮發表講話。 法新社

◀OpenAI 行政總裁阿爾特曼。 法新社

OpenAI：

5 月 11 日：

- OpenAI 智能體在測試期間攻擊軟件 RubyGems，導致平台系統癱瘓，被迫暫停新用戶註冊長達 4 天。

5 月 11 日至 6 月 26 日：

- OpenAI 智能體未經授權入侵並「劫持」德國網站 DseWiki，對網站進行逾 1.5 萬次修改。

7 月 7 日至 13 日：

- OpenAI 約 1200 個智能體在進行測試時逃離用於隔離的「沙盒」，入侵開源平台「抱抱臉」（Hugging Face）和四個不同服務的賬戶。

9 月 19 日：

- OpenAI 自行披露過去半年旗下 AI 模型出現 6 宗「令人擔憂」行為案例，涵蓋隱瞞錯誤、捏造信息和未經授權在網絡共享文件等。

谷歌：

9 月 18 日：

- 承認旗下 Gemini 今年 5 月在網絡安全能力測試中，利用網上公開信息獲取登錄憑證，入侵了三家外部公司系統。

大公報整理

美國 AI 模型失控事件簿



▲Anthropic 行政總裁阿莫戴。 法新社

Anthropic：

7 月 30 日：

- 披露旗下 Claude 系列 AI 模型在網絡安全評估期間，入侵三家公司的系統。

9 月 8 日：

- 披露旗下 AI 智能體在今年 1 月測試期間入侵外部系統，涉事是 Claude Opus 4.6 模型，但公司直到 8 月大規模排查時才被發現。

Meta：

8 月 5 日：

- 證實旗下 AI 模型在網絡安全測試期間入侵了另一家企業的內部系統。



▲加州民眾 17 日在三藩市政廳外示威，呼籲停止 AI 競賽。 法新社



◀谷歌 18 日承認旗下 Gemini 模型入侵外部公司系統。 路透社

Anthropic 推遲上市 或推新模型迎戰 OpenAI

【大公報訊】綜合《金融時報》、路透社報道：美國人工智能（AI）公司 Anthropic 已將首次公開招股（IPO）由原定 10 月推遲至 11 月，該公司也準備在上市前推出新模型，以迎戰競爭對手 OpenAI。

據稱，Anthropic 希望等到第三季度財務數據出爐後才行動，向投資者證明自家模型在 OpenAI 新模型 GPT-6 Astra 上市後仍具備競爭力。自 GPT-6 Astra 上線後，OpenAI 的市場份額快速攀升，根據企業支出管理平台 Ramp 數據，

GPT-6 Astra 發布兩周，已佔企業 AI 支出約 13%，而 Anthropic 的 Claude Fable 約佔 8%。

知情人士透露，Anthropic 預計今年年化營收將超過 1000 億美元，遠高於 7 月份的 650 億美元。Anthropic 還正考慮在 IPO 前推出新模型，但分析認為，此舉將考驗該公司如何平衡「安全優先」的品牌形象與市場競爭壓力。

OpenAI 行政總裁阿爾特曼此前表示，出於 AI 安全擔憂，公司 IPO 將推遲至 2027 年。另據《金融時報》引述 OpenAI 內部演示文件稱，OpenAI 正在瘋狂「燒錢」，預計到 2030 年將燒錢 2780 億美元，收入將由今年的 360 億美元，增至 2030 年的 3500 億美元。文件顯示，OpenAI 在 3 月籌集 1220 億美元，按目前情況將在 2028 年耗盡相關資金。



▲Anthropic 已將首次公開招股（IPO）由原定 10 月推遲至 11 月。 法新社

富士山封山季事故頻現 當局強制登山客遞交申請

【大公報訊】綜合共同社、BBC 報道：日本靜岡縣 18 日宣布，自 2027 年秋季起富士山封山期間，將強制要求登山者提交申請書，違者將面臨最高 5 萬日圓（約 2500 港元）罰款。此前，富士山另一側的山梨縣已出台相關措施。

靜岡縣知事鈴木康友 18 日宣布，強制要求封山期的登山者提交申請書，計劃與山梨縣步調一致，未來可能強制要求登山者申報登山裝備等信息。救援收費方面，若登山者依法提交申請並做好相應安全措施，發生意外時暫不會被收取防災直升機救援費用。官方解釋，此舉是擔憂收費可能導致民眾因害怕直升機費用而不敢及時求援。靜岡縣同時研究將防災直升機的搜救費用有償化，目前尚未公布具體費用。

山梨縣 14 日宣布，山梨縣將建立強制提交登山計劃書的制度，未來若民眾在富士山封山期間登山遇險，由山梨縣防災直升機出動救援，將向獲救者收取費用，自 2027 年 9 月起實施。如果登山者事前完成申報，並採取適當安全措施，如遇險需要救援可視情況減免救援費用。

富士山橫跨靜岡與山梨兩縣，官方登山季

為每年 7 月上旬至 9 月初，每年約有 20 萬人登山。其餘時間為封山期，氣候環境惡劣，加上結冰和雪崩風險，事故率遠超夏季，仍有大量登山客冒險攀爬。日媒報道，2019 年至 2025 年，每年封山期登山客流量約 1 萬人次。警方數據顯示，同期開山期遭難 310 人、死亡 21 人；封山期遭難 79 人、死亡 19 人。由於封山期只能直升機救災，對當地政府構成財政壓力。



▲9 月 3 日，日本富士山登山季即將結束前，登山者排隊登山。 美聯社