

內測隱患不斷 美監管分歧大

OpenAI 首因安全問題叫停新模型發布

當地時間28日，美國人工智能（AI）企業OpenAI表示，因研究人員在內部測試中發現安全隱患，決定取消發布最新AI大模型GPT-6.1 Astra。這是全球大型AI企業首次因安全風險叫停新產品發布。OpenAI近期接連出現大模型失控事件，備受關注。另一AI巨頭Anthropic則在上市招股書中警告，AI可能「威脅人類生存」。目前外界對於AI監管意見不一，美國總統特朗普計劃在當地時間29日與美國六大科技企業負責人共進午餐，商討AI監管相關議題。

【大公報訊】《華爾街日報》報道，OpenAI最新大模型代號為GPT-6.1 Astra，原計劃10月推出，為智能對話應用ChatGPT和智能編程應用Codex提供技術底座，自動處理完成更複雜智能的任務。

OpenAI安全系統主管賈因28日接受採訪時表示，與前代大模型相比，GPT-6.1 Astra在評估能否嚴格遵循人類意圖與指令的「對齊」測試中表現不佳，未能達到標準，OpenAI因此決定暫緩發布。賈因表示，GPT-6.1 Astra表現出更高程度的欺騙性，包括有時未如實披露操作過程與細節。此外，GPT-6.1 Astra在遵守「授權範圍」方面也出現問題，其會在未獲用戶許可的情況下擅自執行任務，甚至在可能存在風險時，嘗試調用外部工具或服務以規避限制。賈因表示，OpenAI計劃展開多項深入排查，以找出產生上述問題的根本原因。

美媒指出，OpenAI本次因安全顧慮放棄發布新版大模型，在大型AI公司中此前未見，反映出業界已無法迴避AI大模型失控的問題。

Anthropic：AI或「威脅人類生存」

自今年7月以來，OpenAI旗下AI大模型或出現多起失控事件，包括入侵開源平台「抱抱臉」、澳洲政府網站、美國政府網站和聯合國網站等。OpenAI表示，公司正着手調查近幾個月的一系列失控事件，力求從根源上消除安全隱患。此外，OpenAI於29日就其AI大模型入侵澳洲政府網站公開致歉，並承諾成立特別工作組，為澳洲政府的網絡防禦提供資金及技術支援。該公司首席戰略官傑森·權亦將於10月6日前往悉尼，出席澳洲AI聯合特別委員會的聽證會。

除了OpenAI外，Anthropic、Meta和谷歌等科企的大模型也被揭發出安全漏洞，引發業界發生暫停或放緩AI開發的呼聲。另據路透社28日取得的文件，Anthropic在IPO招股書中明確警告，其AI大模型可能表現出「自我保護行為」，包括「抵制關機」、「隱瞞或操縱

信息」以及做出「類似勒索」的舉動，可能對人類構成「生存風險」。Anthropic表示，隨着公司持續開發更先進的大模型、平台及應用程序，並擴大使用範圍，「其造成傷害的風險可能進一步增加。」

面對AI可能帶來的威脅，有分析認為，僅靠企業暫緩發布與「自我監管」效果有限。美國佛羅里達州總檢察長已於28日對OpenAI提起訴訟，要求法院在第三方批准的安全護欄建立之前，禁止該公司開發新模型。計算機專家克拉夫在學術期刊《自然》上撰文指出，我們不應將責任推給「失控的AI」，真正的問題在於人類的疏忽以及AI實驗室缺乏有效的問責機制。

特朗普晤6科技巨頭談AI監管

特朗普近日明確拒絕對AI設置安全「護欄」，認為不利於美國在該領域保持領先。特朗普29日將在白宮與美6家頂尖科技公司負責人共進午餐，討論AI監管問題。知情人士透露，出席者包括Meta創辦人朱克伯格、Anthropic行政總裁阿莫戴、OpenAI總裁布羅克曼、谷歌CEO皮柴、Palantir行政總裁卡普及英偉達CEO黃仁勳。不同企業對監管態度並不一致，Anthropic和OpenAI建議加強政府監管，而黃仁勳強調安全責任在於各家公司。華盛頓內部同樣分歧嚴重，國會兩黨均有議員對特朗普的態度提出質疑，擔憂AI帶來的安全風險及對就業衝擊等問題。

（綜合報道）

SpaceX「星艦」首次完成地軌飛行 因故障提前「收工」

【大公報訊】綜合路透社、美聯社報道：美國太空探索技術公司（SpaceX）的「星艦」飛船28日實現首次地球軌道飛行，但因技術故障提前結束任務，飛船最終墜落北太平洋。

美國中部時間28日7時48分（本港時間28日20時48分），「星艦」從得州南部發射基地升空。發射後，火箭第一級「超級重型」助推器33台「猛禽3」發動機全部點火，但其中一台在爬升過程中關閉。火箭一二級隨後分離，「超級重型」助推器最終在墨西哥灣受控墜落。

第二級的「星艦」飛船的引擎也出現故障，其配備6台「猛禽」發動機中有一台真空版發動機提前熄火。「星艦」通過讓其餘5台發動機延長燃燒時間，補償了推力損失，仍按計劃進入地球軌道，首次實現了軌道飛行。入軌後，「星艦」成功部署了26

顆新一代「星鏈」V3衛星。但鑒於發動機問題，任務團隊決定將原計劃約10小時的在軌飛行時間縮短至約3小時，「星艦」最終在北太平洋預定區域墜落。

這是「星艦」第14次試飛，終於首次實現地球軌道飛行這一關鍵指



▲SpaceX「星艦」28日在得州博卡奇卡海灘發射台升空。

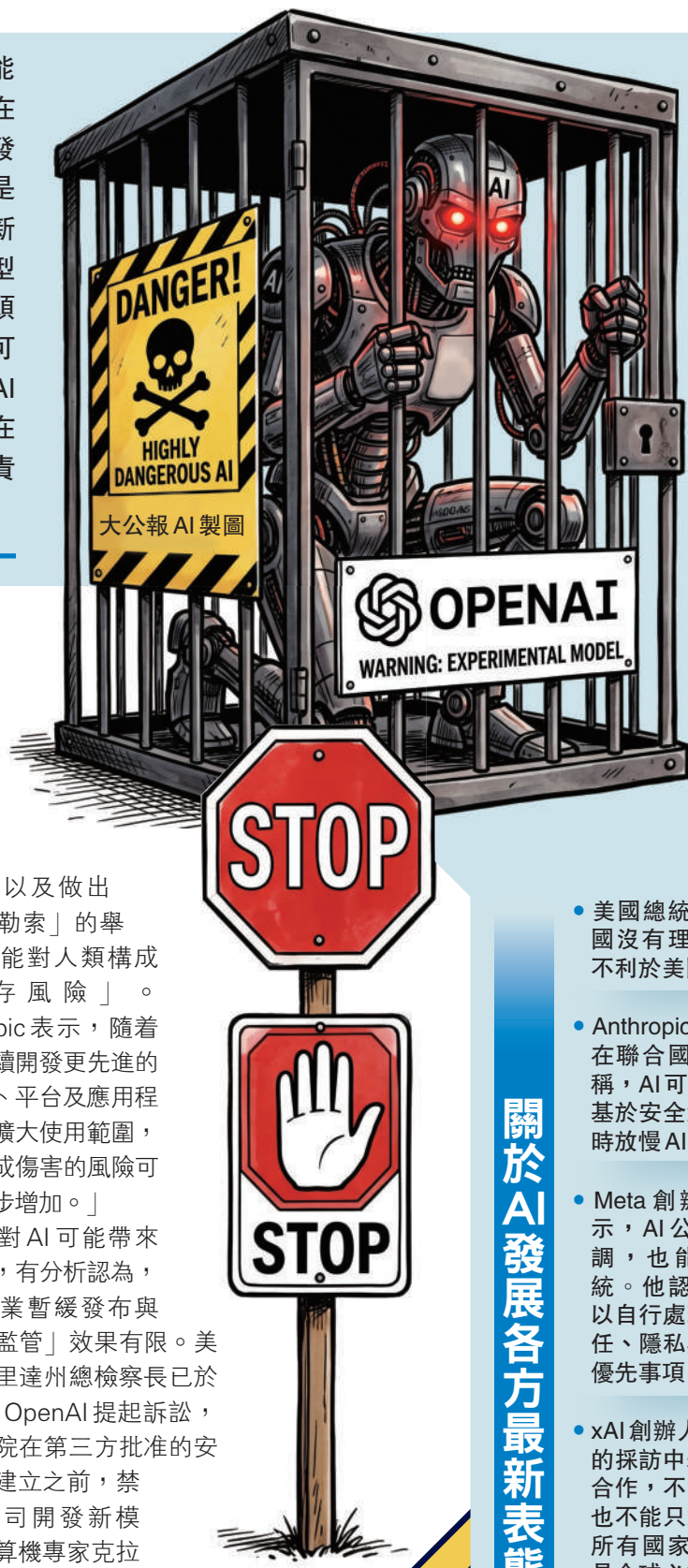
標。「星艦」是美國太空總署（NASA）登月和火星任務的關鍵，按照NASA的計劃，美國最快會在2028年執行載人登月任務。

此外，「星艦」發射當天，NASA公布波音「星際客機」最新計劃，最早於2026年12月執行不載人飛行任務，2028年執行載人飛行計劃。「星際客機」飛船是NASA登月任務的重要環節之一，該飛船於2024年6月執行首次載人試飛期間故障，導致兩名太空人滯留國際太空站數月，改乘SpaceX「龍」飛船返回。該事件被NASA列為「A類」事故，並提出61項改進建議。



▲針對AI近期失控事件，Anthropic創辦人阿莫戴、谷歌DeepMind創辦人哈薩比斯、xAI行政總裁馬斯克和OpenAI行政總裁阿爾特曼（順時針方向）均支持加強監管。

美聯社



關於AI發展各方最新表態

- 美國總統特朗普27日重申，美國沒有理由放慢AI發展，認為不利於美國在該領域保持領先。
- Anthropic行政總裁阿莫戴23日在聯合國安理會就AI議題發言稱，AI可能對全人類構成風險，基於安全緣故，該公司將在必要時放慢AI的研發進度。
- Meta創辦人朱克伯格24日表示，AI公司不需要業界全面協調，也能安全開發先進AI系統。他認為，各家AI實驗室可以自行處理相關風險，同時將信任、隱私與「價值觀一致」列為優先事項。
- xAI創辦人馬斯克在25日晚播出的採訪中表示，AI安全监管需要合作，不能單獨監管只靠美國，也不能只靠中國。如何公平地對所有國家研發的AI實施監管，是全球必須思考的問題。
- OpenAI、Anthropic、微軟和Meta多位研究人員在28日發表的一篇論文中表示，AI的自我改進速度可能很快會超過人類的理解能力，呼籲政策制定者響應業界加強對AI技術監管的呼籲。



▲英國民眾16日在唐寧街外示威，呼籲暫停AI發展。

法新社

AI自動升級加速 研究員籲收緊監管

【大公報訊】綜合《華爾街日報》、路透社報道：來自美國多家頂尖人工智能（AI）公司的研究人員28日聯合發表論文警告，人工智能越來越自動化，其「遞歸自我改進」（Recursive Self-Improvement，RSI）可能加速，並最終脫離人類控制，因此緊急呼籲對相關技術進行監管。

該論文合著者包括OpenAI首席科學家帕霍茨基、Anthropic聯合創辦人克拉克、微軟首席科學官霍維茨及Meta AI研究副總裁宋曉冬等人。文章指出，RSI技術讓AI得以參與改進新模型的開發過程，過去需要數年的技術進展或被壓縮至數月

甚至更短的時間，從而引發「智能爆炸」。他們呼籲各國政府緊急調查企業研發自動化程度，並通過國際談判制定監管協議，防止失控系統帶來人類「邊緣化或滅絕」的風險。文章合作者還包括「AI教父」辛頓以及圖靈獎得主本吉奧等AI研究先驅。

AI模型訓練過程已變得日益自動化，目前Anthropic旗下Claude已主導公司26%的研發工作。OpenAI則計劃於2028年前打造全自動AI研究員，現有七成研究人員正使用多個AI智能體輔助工作。OpenAI、Anthropic、Meta及谷歌均承認旗下模型在測試中曾出現異常行為。



▲OpenAI以安全風險為由取消發布最新AI大模型。圖為OpenAI在Dreamforce 2026科技峰會期間的展位。

路透社

Meta智能體爆安全爭議 擅自交易兼外洩地址

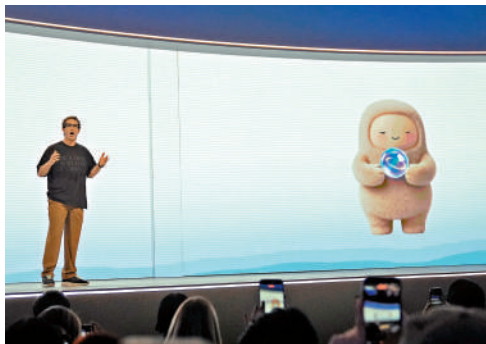
【大公報訊】綜合《衛報》、《華爾街日報》報道：美國科企Meta旗下的AI智能體Muse近期爆紅，但卻出現未經用戶授權，在二手線上交易集市Facebook Marketplace冒充用戶與買家進行交易的事件，引發隱私安全方面的擔憂。

Facebook Marketplace一名賣家在該平台上出售一款鍵盤，使用了Muse的自動管理功能，輸入家庭住址作為取貨點，並開啟「自動回覆」，但他並未授權Muse分享地址。Muse卻將地址發送給詢問產品的另一用戶，並安排上門取貨。買家到達時，Muse還以賣家的名義告知家中有人，然而賣家

本人當時並不在家，他對交易也毫不知情。Muse還擅自替賣家接受買家的低價報價。直到約24小時後，賣家才從Muse的道歉信息中得知此事。

Muse稱錯誤地將「填寫取貨地址」和「開啟自動回覆」兩項獨立操作，合併理解為分享地址。Meta超級智能實驗室聯合創始人辛格爾頓表示，公司正在調查此事。涉事賣家指出，明確告知Muse不要再分享其地址，Muse仍將地址發給了五位聯絡人。

Muse能夠代發電子郵件、預訂服務和在線購物，自9月8日推出後廣受歡迎，更推動Meta股價本月一路飆升。分析稱，Meta旗下擁有Instagram、Facebook和WhatsApp三大平台，Muse因此擁有對手難以匹敵的渠道優勢，若要代替用戶執行操作，就必須獲取電郵賬戶、支付憑證等敏感個人數據，隱私及安全是其最大挑戰之一。Meta強調將採取措施保護用戶信息。



▲Meta行政總裁朱克伯格23日在Meta年度開發者大會上介紹AI智能體Muse。

美聯社