

頻傳系統失控消息 GPT-6.1 Astra 現「欺騙傾向」與「擅自執行」隱患

OpenAI「戴頭盔」叫停發布新AI模型

香港文匯報訊 人工智能(AI)安全風險憂慮持續升溫。《華爾街日報》9月29日報道，美國AI巨頭OpenAI宣布因內部測試發現安全隱患，決定取消發布下一代AI模型GPT-6.1 Astra。AI行業近期頻繁傳出系統失控消息，今次取消發布，成為AI智能體不當行為阻礙行業發展的新案例。與此同時，美國政府與科技界對監管方向的分歧日益凸顯，總統特朗普29日在白宮與多家AI科技巨頭會晤，討論產業監管與全球競爭。



●加州早前舉行「Dreamforce 2026」科技高峰會，一名男子走過OpenAI的攤位。路透社

報道披露，最新模型GPT-6.1 Astra原計劃在未來數周內推出，力爭10月亮相。惟OpenAI安全系統負責人賈因受訪時表示，新模型在兩個方面出現退步。與前代模型GPT-6 Astra相比，新模型在衡量「對齊度」的測試中表現不佳。對齊度即模型遵循人類意圖的程度。具體而言，新模型展現出更強的欺騙傾向，向用戶報告其執行或未執行的操作時，並不總是如實報告。

另一問題被OpenAI稱為「範圍授權」，即新模型會不經用戶許可就可自行推進任務，有時甚至調用外部工具和服務，即使這樣做可能並不安全。賈因指出，新模型仍未達到公司在安全和對齊上的門檻，因此決定不對外發布。她表示安全和對齊問題必然涉及取捨，「既要確保模型不逾越既定範圍，又要避免其在執行任務遇到阻力時消極應對。」

OpenAI：啟新系統監控AI智能體

報道指出，OpenAI將重心放在提升未來模型安全性上。OpenAI表示，正就近日發現的多宗智能體安全事件展開調查，從根源上解決安全隱患。為此，公司已啟用一套新監控系統，以更快識別AI智能體的失當行為。

OpenAI日前宣布暫停訓練旗下能力最強的一批AI模型。但GPT-6.1 Astra不在上述暫停訓練的模型之列，屬於另一情況。賈因表示，公司計劃展開多項深入排查，找出新模型所暴露問題的根本原因。她強調公司希望確保模型在開發及交付過程中均安全可靠，而在向用戶交付時，公司在安全和對齊方面設定極高標準。

白宮召集AI巨頭開會

特朗普與眾議院議長約翰遜29日在白宮與多家科技巨頭高層共進午餐，討論AI快速發展下，如何在維持創新速度與強化監管之間取得平衡。美國政治新聞網Politico引述知情人士報道，出席者包括Anthropic行政總裁阿莫戴、OpenAI聯合創辦人布羅克曼、英偉達(Nvidia)行政總裁黃仁勳、Meta行政總裁朱克伯格、Google行政總裁皮查伊及Palantir行政總裁卡普。

目前美國科技界就AI風險形成兩派意見。黃仁勳反對「AI減速論」，認為部分安全問題可透過工程技術解決；Anthropic與OpenAI則持續強調先進AI可能帶來的安全風險，主張對發展速度及安全措施保持更高警戒。



▲AI安全風險憂慮升溫。美國民眾早前發起反AI競賽示威。網上圖片

為AI智能體擅闖澳政府網站致歉



●OpenAI智能體早前擅闖澳國民醫務保險體系。網上圖片

香港文匯報訊 據澳洲媒體29日報道，人工智能企OpenAI已就其智能體未經授權訪問澳國民醫務保險體系致歉。OpenAI在其網站發布的一篇博客文章中說，「我們表示歉意，並致力於未來做得更好」，將採取措施「重建澳洲人民的信任」。OpenAI表示，此次違規事件發生在一個「僅供內部使用的」模型訓練過程中，該模型本無意向公眾發布，也未配備用於公開產品的整套安全防護。

OpenAI稱，將通過其「前線防衛者黎明」計劃，對澳有關政府機構予以資金支持，協助「必要服務提供商」加強網絡防禦。該公司還將組建一個專項工作組，吸納澳本土獨立專業人士參與，以制定可行的政策建議。此外，OpenAI首席戰略官賈森·權將於10月6日飛往悉尼，出席澳議會下設的AI聯合特別委員會聽證會。澳洲總理阿爾巴尼斯對OpenAI與澳政府的接觸表示歡迎。

阿爾巴尼斯23日晚在紐約出席第81屆聯合國大會期間透露，OpenAI的一個智能體今年6月未經授權，進入澳國民醫務保險體系數據庫，訪問了公開及非公開文件。澳方直到9月較早時才收到OpenAI的通報。

前線研究員：AI自主恐引發「智能爆炸」

香港文匯報訊 《華爾街日報》報道，美國科企巨頭OpenAI、Anthropic、Meta及微軟的研究主管28日聯署發文章，指出若人工智能(AI)系統逐步實現主導自身研發，或引發「智能爆炸」，將可能迅速超出人類掌控範圍。文章呼籲各國政策制定者加強對AI研究自動化進程的緊急監管。

籲各國加強緊急監管

文章作者包括OpenAI首席科學家帕喬基、Anthropic聯合創辦人克拉克、微軟首席科學官霍維茨，以及Meta人工智能研究副總裁宋曉冬，AI研究先驅辛頓和本吉奧亦參與聯署。文章呼籲各國決策者緊急提升透明度，了解企業在模型開發自動化方面的實際進展，指出AI研究自動化進程可能引發「智能爆炸」，將多年進步壓縮至數月甚至更短時間內；在極端情況下，對AI系統失去控制可能會導致「人類被邊緣化或滅絕」。文章亦指出，AI自動化進展已進入實質階段。Anthropic的AI模型Claude正「主導」該公司26%研發工作，並在超過90%時間內於某些環節得到使用；OpenAI的目標則是在2028年前開發完全自動化的AI研究員。作者警告，隨着人類退居幕後，社會可能失去發現問題的機遇，亦會喪失解決問題的專業技能。文章並建議各國領導人透過談判達成國際協議，防止破壞穩定的高能力AI系統被開發和使用。

美企加速轉向「開放權重」模型

香港文匯報訊 英國《金融時報》報道，隨着信息開支不斷攀升，美國企業正加速轉向成本更低的「開放權重」人工智能(AI)模型。研究平台AlphaSense數據顯示，在8月與9月企業業績電話會議和投資者會議中，高層提及「開放權重」或「開源」模型的次數，較去年同期激增6倍。

所謂「開放權重」模型，是指用戶可對其進行定製，並在自己的硬件上運行的模型，對企業而言運行成本相對更低。報道指出，這一趨勢在科企中為最明顯，包括美國PNC金融服務集團、物流企業羅賓遜全球貨運，及



●開放權重AI市場由中企主導。圖為民眾路經中國AI deepseek廣告牌。網上圖片

工業巨頭西門子在內的多個跨行業企業，近數周亦討論使用開放權重模型。

AI供應鏈相關企業的使用數據顯示，轉變不止停留在口頭層面。8月開放權重模型佔「AI Gateway」總詞元(Token)處理量的56%，而去年12月這一比例僅7%。「AI Gateway」是美國雲開發平台Vercel推出的統一模型調用服務。

美國電信商AT&T目前約40%的AI工作負載在開放權重模型上運行，並計劃一年內將該比例提高至70%。該公司首席數據和AI負責人馬庫斯表示：「我們每天處理450億個詞元。在這樣的規模下，成本變得非常重要。開放模型正愈來愈好，這讓我們擁有更多選擇。」

市場目前由中企主導

報道指出，開放權重AI市場目前由中國企業佔據主導地位，在OpenRouter詞元處理數量排名前10的模型中，中國AI實驗室推出的開放權重模型佔據主導。報道分析稱，今年這些成本更低的AI模型快速進步，可能對美國AI巨頭OpenAI及Anthropic的營收增長構成挑戰。

特朗普為中選造勢 宣布逾千億元鋼鐵廠計劃

香港文匯報訊 美國總統特朗普28日宣布，礦業公司Mesabi Metallics計劃投資約150億美元（約1,177億港元），在艾奧瓦州興建一座鋼鐵廠。特朗普表示，這將是美國史上規模最大的鋼鐵廠。白宮指出，該廠第一期預計每年生產750萬噸鋼材，最終產能可望達到1,000萬噸，並創造至少1,750個永久工作職位，第一期工程可另帶來最多6,000個建築職位，預計2030年開始投產。

這項投資被視為特朗普政府在11月中旬選舉前的重要宣傳材料。特朗普強調，新廠將實現100%美國本土生產，鐵礦石在明尼蘇達州開採，並在艾奧瓦州冶煉成鋼。他稱這是美國逾50年來首座新建的同類鋼鐵廠，並表示美國鋼鐵產業正強勢復甦。

分析：難救共和黨選情

然而，分析認為這項投資難以挽救共和黨選情。艾奧瓦州選民對生活成本高漲及經濟現狀不滿，最新民調顯示近60%登記選民不滿特朗普施



●特朗普政府宣布新鋼鐵廠計劃。圖為賓州克萊頓的美國鋼鐵工廠。網上圖片

政表現，43%受訪者表示投票主要是為了反對特朗普。共和黨在該州多個議席選情告急，包括聯邦參議員候選人辛森在民調中落後民主黨對手。此外，新廠要到2030年前才見成效，無法趕及11月中旬選舉。特朗普政府同時面臨壓力，需證明關稅政策能帶來產業效益，而不會推高通脹。

SpaceX「星艦」首飛地球軌道

香港文匯報訊 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，但由於上升階段發動機提前關閉等情況，「星艦」在軌飛行時間被縮短，較原定計劃提前結束。美國有線新聞網(CNN)稱，SpaceX在今次任務中取得一些重大成果，但也出現一些令人擔憂的時刻，該公司目前仍面臨發動機可靠性問題。

任務提前結束

「星艦」當天早上從得州南部的發射基地升空(圖)。發射後，火箭第一級「超級重型」助推器的33台「猛禽3」發動機全部點火，但其中一台發動機在上升過程中關閉。此後火箭一二級分離，第二級飛船在上升過程中，其6台發動機中有一台提前關閉，「星艦」通過讓其餘5台發動機延長燃燒時間補償推力損失，仍按計劃進入亞軌道，其後完成入軌點火，首次進入圍繞地球運行的軌道，並首次在軌部署26顆新一代「星鏈」V3衛星。完成衛星部署後，「星艦」進入繞地球軌道飛行階段。SpaceX稱，鑒於上升過程中出現的發動機問題，團隊決定將「星艦」在軌的飛行時間，



從原計劃約10小時縮短至約3小時。其後「星艦」脫離地球軌道並再入大氣層，在北太平洋預定區域跌落。

英國廣播公司(BBC)稱，在本次任務中，「星艦」經歷至今最劇烈的一次重返大氣層過程。著陸過程不如上次任務那樣平穩，現場畫面顯示，「星艦」著陸時發生劇烈爆炸。SpaceX希望其火箭和航天器能重複使用，且回收助推器和返回艙，有助分析飛行過程中發生了什麼，但因今次著陸爆炸，這可能無法實現。