

GPT-6.1 Astra 測試揭發 具擅自執行及欺騙傾向 存安全隱患 OpenAI 停發新模型

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。另外，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。

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，並不總是如實相告。

白宮宴請科企高層商AI發展

另一問題被OpenAI稱為「範圍授權」，即新模型會不經用戶許可就自行推進任務，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



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。」

報道指出，OpenAI將重心放在提升未來模型安全性上。OpenAI表示，正就近月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。

與此同時，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」；而Anthropic阿莫戴與OpenAI阿爾特曼等則主張對發展速度及安全措施保持更高警戒。

美科企巨頭聯署文章促監管AI

另外，《華爾街日報》29日報道，美國科企巨頭OpenAI、Meta及微軟等研究主管聯署發表文章，指出若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，或導致「人類被邊緣化或滅絕」。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。

伊官員對美伊談判前景感悲觀



■伊朗民眾在納傑夫阿什拉夫國際機場外示威，要求放寬制裁，重新開放機場。 法新社

美國與伊朗的談判陷入僵局之際，彭博社28日報道，伊朗官員私下對重開霍爾木茲海峽及結束美伊戰爭的談判前景感到悲觀，預料不會在美國中期選舉前，與特朗普政府達成協議。消息人士稱，美伊雙方上周在紐約聯合國大會期間舉行的會談成果甚微，伊方甚至認為雙方在中期選舉後衝突升級的可能性很高。

報道指出，伊朗官員表示願意尋求外交解決方案，但不相信特朗普政府會遵守協議。美方同樣不信任伊朗，美媒此前更指特朗普告知幕僚，計劃在11月中期選舉後恢復對伊朗的轟炸。一名匿名美官員表示，除非有關伊朗核問題得到解決，否則美方不會與伊方達成協議。

特朗普否認解除制裁換核讓步

美國總統特朗普28日發文，否認美媒有關其願意以解除對伊制裁並解凍資金、以換取伊朗在核問題上讓步的報道。他指該報道「不實」，要求有關媒體立即撤回這篇「虛假報道」。但特朗普向媒體證實，美伊談判代表已透過調解方交換信息。

SpaceX「星艦」首實現地球軌道飛行

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，任務團隊決定縮短「星艦」在軌飛行時間，較原定計劃提前結束。

在當地時間早上7時48分，「星艦」從位於得州南部的發射基地升空。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。「超級重型」助推器隨後在墨西哥灣附近海域受控墜落。

在軌部署26顆「星鏈」V3衛星

第二級「星艦」飛船在上升過程中，其6台「猛禽」發動機中有一台真空版發動機提前關閉。「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。其後任務團隊對「星艦」進行評估，確認後續操作所需的3台海平面版「猛禽」發動機狀態良好，決定繼續執行入軌操作。「星艦」完成入軌點火，首次進入繞地球運行的軌道。隨後「星艦」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表示，鑒於上升過程中出現的發動機問題，團隊決定縮短「星艦」在軌飛行時間，從原計劃約10小時縮短至約3小時。此後「星艦」執行首次離軌點火，脫離地球軌道並再入大氣層，在北太平洋預定區域墜落。



■SpaceX公司的「星艦」火箭從得州基地進行試飛。 美聯社

企業創新指數新里程



南窗集

羅文華

銀行專業會計師、公司董事、香港銀行學會考試評審委員會委員

香港特區政府本年9月發表的《第一個五年規劃(2026-2030)》提出「加快創新科技產業發展」。創新，是這個時代的主旋律，亦是香港未來發展的大方向。

香港中文大學亞太工商研究所(APIB)編制的「企業創新指數」(CII)已踏進第三階段。多年來為香港工商業的創新作出貢獻，參與的大企業越來越多。

筆者在本年7月15日出席APIB聯同香港總商會舉辦的「企業創新指數3.0——以策略基準及最佳實務推進香港的商業創新」研討會，獲益不淺。APIB總幹事張惠民教授以「測試矩陣：從內部看CII3試驗」(Testing the Metrics: An Inside Look at the CII 3 Pilot)為題簡介CII的最新發展：以CII1為基礎，建立以香港為中心的概念架構及啟動評估矩陣改進企業創新；CII2為聚焦於中小企創新能力及資源優化；CII3是用三年的縱向數

據，作探討策略基準及最佳實務做法之用。

張教授繼而指出創新架構的工具精益求精。其一是數字化，由數字用途，演變至建立數字基礎及數字化，以及AI轉型及管理；其二是人才管理，由員工培訓及發展，演變至高級才幹之充分使用；其三是可持續性的良性影響，演進至融合新項目。

最後，張教授指出，今次CII3的問卷，有幾處改良的地方。其一是預先填入已有的資料，即使用大模型從上市公司的年報中擷取相關資料，省卻填表的人工工夫；其二是基於網頁的補充數據，即從公司的網頁中搜集公開資料，作為補充資料之來源；其三是一致性，所有從網頁搜集得來的不同公司的資料均定在同一時點，以示公允。

筆者認為，中大商學院APIB推動香港工商業的種種努力值得嘉許，其播下的創新種子，將會持續發芽、開花、結果，轉變為商業用途，為香港增添光彩。