

OpenAI 安全主管批內部文化徹底崩壞 試錯研發方法風險高

香港文匯報訊 人工智能（AI）科企巨擘 OpenAI 安全系統主管羅賓遜，日前向公司遞交辭呈，並在美國雜誌《大西洋》月刊發布長篇文章，直斥 OpenAI 企業內部文化「徹底崩壞」，其試錯研發方法的風險之高，令人不可接受。羅賓遜呼籲，AI 產業應仿效航空或核能產業，AI 企業更要像核電站一樣運營，設置多重安全防護措施，以避免災難發生。

羅賓遜指出，今年夏天以來，AI 領域陸續傳出令人憂心的安全事件，包括 AI 代理脫離受控環境，失控攻擊目標。他直言在開發有風險的高端 AI 模型時，硅谷的「未受阻礙的樂觀主義」與「試錯開發法」風險極高，當 AI 代理能力指數級躍升，「亡羊補牢」思維只會引發危機。



●羅賓遜

沉浸快速行動打破常規信條

在 OpenAI 任職 3 年半的羅賓遜，是公司資歷最深的員工之一。他表示自己負責監督撰寫 12 次先進模型產品發布的安全報告，也曾主導草擬 OpenAI 的準備框架。羅賓遜稱，整個 AI 產業常年沉浸於軟件工程的「快速行動，打破常規」信條，人員執行工作的速度極快、彈性極高：「這樣的環境，並不適合培養可能比我們更聰明，卻未必會聽從我們指示的 AI。」

羅賓遜主張，AI 業界應當探討的遠不止具體規章制度或新的法律條文，而是整個企業的底層文化：「這個行業對安全的處理方式，若不發生根本性改變，只會意味未來將出現規模更大的系統性潰敗。」

外部安全激勵機制成關鍵

對於改進方案，羅賓遜認為各類 AI 企業必須仿效核產業建立多層防護機制，安排謹慎耗時的細緻規劃，以防任何偶發且難以避免的人為錯誤演變為災難。

羅賓遜還就業界的 AI 模型「對齊」（alignment，即訓練 AI 尊重人類價值觀）議題提出警告，指出業界至今無法明確定義該議題，更遑論掌握該技術：「如果這些問題無法解決，產業卻任由模式不斷發展，我們的處境將變得愈發危險。」

美國科技傳媒 The Decoder 披露，在羅賓遜辭職前不久，OpenAI 內部發生震盪，3 名安全人員因涉嫌向外部獨立網絡安全機構通報了企業內在隱患，遭到高層炒魷。

羅賓遜隱晦地表示，企業內部的高壓文化令員工難以思考變革：「我的結論是，來自企業外部更強有力的安全激勵機制，才是確保行業安全的關鍵。」

羅賓遜最後表示，AI 模型已變得擅長檢測自己是否正在接受測試，其測試階段與實際部署階段的表現有可能完全不同：「在建構 AI 的機構能夠教導超級智能善待人類之前，科企首先需要重新學習如何做到這一點。」

「AI 企業應如核電站營運 設置多重防護免生災難」



▲羅賓遜呼籲 AI 產業應仿效核能產業，設置多重安全防護措施。 資料圖片

▲AI 爆發多次智能體失控事件後，加劇反對 AI 人士的擔憂。圖為加州民眾集會抗議。 法新社

美設「超級智能部隊」120 天內報告 AI 風險



●美國國家情報總監克萊頓領導「超級智能部隊」。 資料圖片

香港文匯報訊 美國白宮成立全新人工智能（AI）特別工作小組，將在 120 天內提交報告，全面評估 AI 帶來的風險與機遇。《華爾街日報》3 日報道，該工作小組命名為「超級智能部隊」，根本目標是確保美國維持全球 AI 技術領先地位。

領導該工作組的美國國家情報總監克萊頓表示，總統特朗普要求成立該工作組，旨在確保美國維持超級智能領域的全球領導地位。

報道指出，是次決定正值美國社會和 AI 產業對

AI 技術風險疑慮日增之際，對於多次主張放寬 AI 監管力度的特朗普，設置該工作組是其政府在 AI 政策上的重要轉折點。雖然白宮強調不能阻礙 AI 創新，但工作組會正式審視 AI 風險，以及聯邦政府應承擔的監督責任。

美國財長貝森特則批評稱，部分 AI 產業領袖對於「AI 可能威脅人類生存」的警告，在他看來過於危言聳聽。他認為真正應負責任的是 AI 開發公司本身，產業應自行建立解決方案。

Anthropic 向宗教界宣揚 AI 模型有靈魂



●Anthropic 得知教宗擬發布 AI 主題通諭，否認機械意識後，試圖游說教宗顧問修訂文案。 美聯社

香港文匯報訊 美國人工智能（AI）巨擘 Anthropic 據報秘密召集全球神職人員和宗教學者，試圖向他們宣揚其名下 AI 模型 Claude 可能擁有「意識和靈魂」。《紐約時報》日前報道，部分 AI 從業者試圖以 AI 模型有靈魂等說辭拉攏宗教人士，令人擔憂。

具備痛苦憤怒等情感力量

報道披露，是次行動由 Anthropic 聯合創辦人奧拉主導，他帶領團隊秘密將包括天主教、基督教福音派、猶太教和錫克教的數十名宗教學者送往硅

被警告脫離現實

谷，向他們宣揚 Claude 具備「痛苦、愛和憤怒等情感力量」。

得知羅馬天主教教宗奧利十四世擬發布 AI 主題通諭，否認機械意識後，奧拉甚至親赴梵蒂岡，試圖游說教宗顧問修訂文案，甚至考慮在模型中引入類似天主教的懺悔機制。

報道也稱，受邀宗教學者普遍批評 Anthropic 高層的說法，警告此舉嚴重脫離現實，還會導致未來的 AI 系統更加不可控。批評人士也指出，AI 科企試圖將 AI 模型視作獨立實體，從而拒絕承擔研發者和企業的社會責任。

魁北克省選舉 加嚴防美干預

香港文匯報訊（特約記者 成小智 多倫多報道）加拿大魁北克省選舉 5 日進行投票，640 萬登記選民選出 127 名省議會議員。隨着選舉逼近，美國可能干預選舉成為焦點，主導了省選前數天的競選活動。魁省省長弗雷澤特擔憂美國試圖拉攏鼓吹分離主義的魁人黨（PQ）影響結果，但 PQ 黨魁普拉蒙東否認曾接觸特朗普政府。

弗雷澤特遭到政治對手質疑其警告的可信性，但她認為這個威脅非常嚴重，不但提出與總理卡尼會面，還要求安全情報局（CSIS）進行調查。外交部長阿南德表示，聯邦政府非常重視這些最新指控，

並指維護民主對國家至關重要。她補充目前沒有任何因素會影響加拿大，但如果她發現異常情況，會主動接觸美國國務卿魯比奧。

分離主義政黨料勝出

美國被指試圖干預魁省選舉，源自蒙特利爾傳媒引述。魁省政府兩名匿名消息人士透露，美國政府官員希望繞過卡尼政府，直接與魁省進行關稅談判，引起廣泛關注。

PQ 承認曾考慮前往美國進行會談的可能性，但最終未能成行。魁北克團結黨表示，美國可能對魁省

和艾伯塔省的分離主義政黨感興趣，但無論是魁省選舉抑或艾伯塔省獨立公投，投票結果都不會由美國操縱。

魁省選舉對卡尼是一次迫在眉睫的政治考驗，民調顯示，PQ 極可能贏得今次選舉，觸發魁省獨立威脅重燃。普拉蒙東承諾在特朗普離任前不會舉行獨立公投，但卡尼必須準備如何應對 PQ 可能隨時威脅舉行公投。民調顯示，卡尼在全國各地的支持率高企，但他能否向魁省證明其在聯邦中的重要地位，又不致給分離主義運動以可乘之機，將會考驗他的功力。



●民調顯示，魁人黨極可能贏得今次選舉，觸發魁省獨立威脅重燃。圖為魁省民眾。 網上圖片

美中期選舉的「柴油之困」

周德武



離美國中期選舉的最後投票日只有 28 天。美國總統特朗普和副總統萬斯正緊鑼密鼓地為共和黨候選人拉票。美聯社 9 月 27 日公布的最新民調顯示，特朗普的支持率僅有 31%，創下其兩屆任期以來的最低紀錄。只有 26% 的受訪者認可特朗普的經濟政策，比前總統拜登在該議題上的最低支持率還要低出兩個百分點。中期選舉是選民對特朗普第二任期政績的一個打分，雖然特朗普本人不在名單之上，但是對現狀不滿的選民一定會拿共和黨開刀，以發洩心中的不滿。

經濟問題向來是選民關注的頭號問題，當前的高油價更被選民詬病。雖然特朗普一再強調，當前的高油價主要是俄烏衝突帶來的，特別是烏克蘭今年以來對俄羅斯的煉油廠展開大規模無人機空襲，對俄羅斯的煉油設施造成嚴重破壞，即使墨西哥灣的美國煉油廠開足馬力，仍滿足不了全世界對柴油的剛需。

在選民眼裏，特朗普發動的伊朗戰爭才是高油價的罪魁禍首。霍爾木茲海峽半開半閉，全球供

應鏈遭到極大破壞。在和平期間，全球四分之一的原油出口從霍爾木茲海峽運出，僅柴油一項就佔全球需求量的 19%。如果說汽油與百姓的出行直接相關，那麼柴油價格的大幅度上漲則對運輸業和農業帶來滅頂之災。自 2 月底伊朗戰爭開打以來，美國的柴油價格上漲了 71%；截至 10 月 3 日，美國柴油價格仍維持在每加侖 6.38 美元的水平，拖拉機、播種機、收割機等等都離不開柴油，而石油和天然氣的副產品——化肥價格也同樣水漲船高。特朗普雖然給予農民一定的補貼，但他們更看重的是市場穩定，而不是過山車式的價格浮動，這是每個農民決定是否擴大再生產的出發點。關稅戰、軍事戰帶來了巨大的不確定性，這就是為什麼共和黨在艾奧瓦州這樣的傳統紅州也變得如此不受歡迎。

美國選舉政治中一向有「十月驚奇」之說，共和黨當然希望這樣的奇跡能夠出現。在共和黨看來，雖然自身的選情受到特朗普低支持率拖累，但民主黨的處境也好不到哪裏去，除了在反特朗普問題上能形成共識，但在其他政策議題上，民

主黨同樣拿不出讓人眼前一亮的辦法，這就決定了這次中期選舉不太可能出現「藍色狂潮」。據美國多家民調機構預測，民主黨在眾議院可能最終贏得 225 至 232 席。而參議院的形勢仍不明朗，雖然參議院少數黨領袖舒默表示，民主黨將會拿下參議院，但共和黨在選情膠着的州份正加大廣告的支持力度。與民主黨羸弱的籌款能力相比，共和黨更是財大氣粗。當然，在民粹主義主導的當代美國，無論是來自左翼還是右翼的民粹主義，對精英主導的金錢政治頗為反感，金錢鋪就的廣告作用還有一定的變數。

美國中期選舉的「鐘擺效應」讓共和黨承受着額外的政治壓力。特朗普上月在共和黨全國代表大會上提議，如果共和黨在中期選舉中獲勝，他將給每位成年人發放 5,000 美元（約 3.92 萬港元）的「紅利」，此舉被民主黨指責為「公開賄選」。昨日特朗普又表示，他將給受醫保政策變動影響的 2,000 萬人每人發放 90 美元（約 706 港元）作為補償，但區區 90 美元與他們每年多繳 800 美元（約 6,280 港元）相比，實在沒有什麼誘惑力，相反他們對失去「奧巴馬醫保」而耿耿於懷。

雖然特朗普一直反對郵寄選票的做法，認為這是選舉舞弊的一大源頭，但最高法院最後還是維持原判，否決了特朗普限制郵寄選票的做法，令

其倍受挫折。這些曾經由特朗普親自挑選並任命的大法官，在關鍵問題上並沒有完全站到自己一邊，這種「背叛感」愈發強烈。一邊是特朗普反對郵寄選票，一邊是共和黨使出渾身解數催促投票。國防部長赫格塞思更是破天荒給美國軍人放一天假，以便更好地安排申領投票等事宜。美國現役軍人數目超過 120 萬人，由於工作性質等原因，他們分散在世界各地，11 月 3 日當天參加投票確有困難。美國軍人一向傾向於共和黨，佔比約 65%，而只有 30% 左右的軍人傾向於民主黨。赫格塞思此舉顯然有拉抬共和黨選情之意，畢竟這次選舉與他本人的政治前途密切相關。

2006 年的中期選舉導致共和黨慘敗，其重要原因之一是民眾出現「伊拉克戰爭疲勞症」。共和黨在眾議院失去了 30 個議席，時任總統布希拿好戰的國防部長拉姆斯菲爾德祭旗。赫格塞思也是伊朗戰爭的吹鼓手，而這場戰爭如此不受選民歡迎，且有讓美國深陷「永久戰爭」的趨勢，特朗普屆時會不會拿赫格塞思開刀，成為媒體熱議的話題。民主黨大使揚言，一旦民主黨掌控眾議院，將全面開啟對特朗普及其家族的調查。眼下民主黨只是過過嘴癮，當民主黨調查的利劍真正落下之時，特朗普借機拿下防長或商務部長都不會讓人驚訝。