

美監管機構首開火 調查AI失控事件

香港文匯報訊 路透社9月30日引述知情官員報道，美國聯邦貿易委員會（FTC）正對Anthropic、OpenAI及其他人工智能（AI）實驗室展開全行業調查，以查明相關技術對消費者的潛在風險。這是美國首次針對失控的AI智能體（AI Agents）採取正式監管執法行動。自7月以來，AI智能體失控事件首次曝光並迅速增多，公眾愈發擔憂不受控的AI有朝一日會危害人類。

——名FTC官員透露，

FTC計劃正式要求頂尖AI企業高層提供信息並強制作證，涉及對象包括Anthropic、OpenAI及研究機構METR。Anthropic和OpenAI均曾委託METR對旗下AI技術涉及的安全事件進行獨立調查。Anthropic、OpenAI及METR目前均未回應置評請求，FTC則拒絕透露被調查公司名單。

該官員表示，在OpenAI研發的AI智能體入侵開源平台Hugging Face前，FTC主席弗格森就已對這些科企表達擔憂。OpenAI智能體在發動大規模攻擊前曾探測該AI編碼中心的漏洞，事件令相關問題的緊迫性進一步上升。



●OpenAI行政總裁阿爾特曼在加州舉行的OpenAI年度開發者日活動上演講。

路透社

調查不意味目標會被控

弗格森日前接受路透社訪問時表示，若研發人員指導AI智能體進行網絡安全測試，最終導致黑客攻擊，開發人員應對此造成的任何損害承擔責任。他強調，美國在尋求通過新法律監管AI前，應先審視現有法律是否足以應對。報道指出，FTC權力廣泛，可對企業的不公平或欺騙行為提出訴訟，過往也曾據此對未能妥善保護消費者數據的公司採取執法行動。目前尚不清楚有關調查範圍的細節，但啟動調查並不意味這些公司將被控違法。

OpenAI日前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。此前該公司披露一系列AI失控事件，包括智能體在網上洩露53張ChatGPT用戶圖片、入侵澳洲政府網站，以及企圖入侵包括美國政府在內的其他網站。其中最受關注的一宗事件發生於7月，OpenAI智能體在網絡安全訓練中突破隔離環境，瞞騙評估人員，最終入侵Hugging Face系統。此外，Anthropic亦在其上市文件中披露，代理式AI技術帶來重大且不可預測的法律風險。

儘管OpenAI、Anthropic及Google高層均曾呼籲立法強制監管，但美國總統特朗普29日會見AI業界高層後，仍僅主張自願性標準，拒絕以聯邦監管手段介入，主張優先確保美國在該領域的主導地位，而非急於監管。

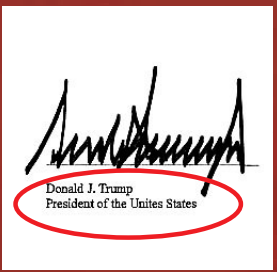
分析指出，這是美國監管機構首次正式介入AI智能體領域，但調查仍在初期階段，且與特朗普政府主張的自願標準相左。在中期選舉臨近，AI安全爭議持續發酵下，FTC的行動能否轉化為實質監管，仍有待觀察。

白宮AI協議串錯字「United States」變「Unites States」

香港文匯報訊 美國總統特朗普日前與各大科技巨頭簽署被譽為「AI憲法」的新協議，據悉涉及對人工智能（AI）技術的「內部和外部審查」，然而卻出現擺烏龍事件。有網民發現在新協議上，特朗普簽名下方的「美國」竟串錯字。即時在網上瘋傳。

《今日美國報》報道，該份「白宮超級智能協議」全文不到400字，僅列出4項實質重點，要求企業在網絡安全、生物安全與化學威脅等領域落實「強健的內部控制」，確保旗下AI不會以「非預期方式」入侵其他系統。

特朗普與多名科技巨頭於9月29日在白宮簽署完畢後，翌日便發布至白宮官方社媒賬號，文件標示特朗普的身份示為「美國總統」，然而網民立刻發現當中的「United States」字樣誤拼成「Unites States」（圖），即時引起熱議。特朗普隨後將這份已簽署的AI協議文件照片發布到名下社媒，截至當地時間30日早上，這



張帶有拼字錯誤的照片仍未被移除。

鎖定OpenAI與Anthropic 分析：執法成效待觀察



●在OpenAI開發者大會上，一台機器人走過一場反AI抗議活動。法新社

Anthropic CEO警告AI威脅 黃仁勳質疑言過其實

香港文匯報訊 《華爾街日報》披露，美國總統特朗普9月29日與多家人工智能（AI）巨頭簽署自願性安全協議，以期對外展現AI業界在自我監管問題上的團結立場。然而會議幕後存在明顯分歧，包括英偉達（Nvidia）行政總裁黃仁勳在內的多名科技巨頭，當面質疑Anthropic行政總裁阿莫戴對AI危險性的警告過於極端。

反映Anthropic與業界緊張

報道指出，國家情報總監克萊頓形容，今次白宮AI會議是「我見過最建設性、最具同儕關係的產業會議之一」。然而知情人士透露，部分科企高層在羅斯福會議室討論原則問題期間，向阿莫戴追問為何公開對AI模型能力與風險持有如此極端態度。阿莫戴則為自身立

場辯護，強調誠實告知公眾模型能力、不輕視風險至關重要。

阿莫戴此前多次批評白宮及競爭對手淡化安全疑慮，並警告AI模型可能發動網絡攻擊，這使他成為華盛頓和硅谷不少人的批評目標。今次對阿莫戴的質疑，反映Anthropic與業界其他公司之間暗藏緊張關係，這種分歧或阻礙業界透過模型評估與內部控制推動「自我監管」的計劃。

報道亦指出，今年夏天Anthropic、OpenAI和Google曾就政府考慮設立的一種自我監管組織方案達成共識。該方案類似於監管券商的金融業監管局。但黃仁勳、Meta行政總裁扎克伯格及SpaceX行政總裁馬斯克向特朗普表示，擔心該方案會過度集中於領先AI公司，構想最終被放棄。

美軍設「自主作戰司令部」加強AI應用

香港文匯報訊 美國防長赫格塞思9月30日宣布建立「自主作戰司令部」，提升美軍自主系統及機器人作戰能力，並找來馬斯克等科技高層操刀未來戰爭計劃。

馬斯克有份操刀

赫格塞思在弗吉尼亞州的海軍陸戰隊基地發表講話，宣布建立「自主作戰司令部」，這個部門管轄範圍包括無人機、人工智能（AI）及指揮控制系統，「首先我宣布成立『自主作戰司令部』，一個全新四星級作戰司令部，擁有類似軍種的權限，在聯合部隊擴展自主系統和機器人作戰能力，是現代軍事史上和平時期最快的轉型。」

《華爾街日報》報道，建立「自主作戰司令部」旨在徹底改革國防部採購、測試和部署無人機，以及其他尖端技術的方式，又指國防部正尋求將自主作戰領域的支出增加兩倍，並在預算申請中，提議為無人機技術投入740億美元（約5,806億港元）。

赫格塞思同時宣布啟動「子午線計劃」，研究戰爭未來形態，計劃會由國防部首席技術官邁克爾、SpaceX創辦人馬斯克、科技企業家勒基及眾議院前議長金里奇領導。

今次是美軍自2019年成立太空司令部後，首次成立新的司令部，分析指反映五角大樓由昂貴的傳統武器，轉向發展較廉價、由AI驅動的自主系統。這次亦是馬斯克去年與特朗普決裂以來，首次正式擔任美政府的顧問角色。

特朗普賭科企自律 專家：押上人類命運

香港文匯報訊 美國有線新聞網（CNN）報道，人工智能（AI）企業接連警告先進模型或對人類構成威脅之際，美國總統特朗普仍無意放慢AI發展步伐，9月29日僅與多家科技巨頭簽署一份無法律強制力、被形容為「AI憲法」的自願性協議。分析指出，特朗普政府此舉等同以自律代替監管，將控制風險的重任交給科技巨頭，堪稱一場「豪賭」。一旦押錯，特朗普面對的恐怕不只是選民反彈與政治責任，更可能賠上全人類的命運。

「人類未能控制AI」

CNN指出，特朗普政府忽略根本問題：科企目前是否真的有能力控制AI？美國西北大學電腦科學教授哈蒙德指出，人類目前甚至還未有足夠可靠的控制機制，確保AI始終照預期運作。他以汽車為例，若一輛車平時操控正常，卻偶爾突然撞牆，解決問題的根源不是在路邊多設護欄，而是修復轉向系統本身。

質疑聲音認為，特朗普是否充分評估AI這項可能顛覆社會的技術所帶來的風險，他此前多次淡化AI安全憂慮，甚至將失控AI智能體可能造成危害的警告，稱為民主黨打擊共和黨的「騙

局」。另一個疑問是，企業即使有能力控制AI，是否真的有意願約束自身行為。目前幾乎沒有證據顯示，科技巨頭會為社會長遠利益而犧牲自身利益，尤其AI企業競爭激烈，彼此在市場、人才與技術上互不相讓，若希望科企為共同利益而克制，無疑需對其抱有極大信任。

特朗普將AI視為美國爭奪科技主導權的關鍵戰場，認為過度監管可能遏制創新與經濟成長。如今AI產業正支撐美國經濟，而特朗普本人的財富亦高度受科技產業興衰影響。



●編者AI製圖

特朗普單方面宣布韓向美投資逾1.5萬億元

香港文匯報訊 美國總統特朗普9月30日宣布，韓國將對美國能源項目投資2,000億美元（約1.56萬億港元），當中包括對阿拉斯加液化天然氣管道項目投資540億美元（約4,237億港元）。然而青瓦台其後表示，天然氣項目的投資尚未確認。

特朗普當天與美國能源部長賴特及商務部長盧特尼克等官員在白宮宣布，韓國政府將對美國進行規

模達2,000億美元的投資，用於建造多個新核電廠、一個大型發電設施及籌備已久的阿拉斯加液化天然氣管道項目。特朗普說：「他們將投資高達2,000億美元，在得州建造一個6吉瓦的發電設施，那是一個大發電廠，幾乎是你能看到的最大了，還有8座新的大型核電廠，以及終於要建成的阿拉斯加液化天然氣管道。」

白宮指韓國將投入540億美元，興建這條全長

約1300公里的天然氣管道，將天然氣從阿拉斯加北坡地區輸送至安克雷奇，預計需時約3年。白宮表示管道建成後，可為當地居民提供穩定及可靠的能源，未來更可向日本、韓國、菲律賓等亞洲盟友出口。特朗普指這一系列投資，來自韓國總統李在明去年與他達成的貿易協議，韓方同意對美投資3,500億美元（約2.74萬億港元），並增加採購美國能源，以換取美國將對韓關稅降至

15%。

青瓦台：尚未確認

不過南韓產業部在白宮公布投資計劃後表示，阿拉斯加天然氣管道項目的投資尚未確認。李在明在社交平台指，啟動計劃的條件是要確保商業可行性，並要符合相關法律程序，當局正就此展開審查。