美國《華爾街日報》分析指出，人工智能（AI）發展之快、性質之新，已遠超人類社會既往經驗，令各國監管機構面臨前所未有的考驗。從核電、噴氣式飛機到藥物和金融衍生品，人類對高風險技術的監管已有數十年經驗，但這些歷史無法直接套用於AI，原因在於AI既能帶來巨大益處、又極其危險，但卻可自由獲取，而人類知識更可能追不上其發展，沒有任何模板可供各國遵循。 ●香港文匯報記者 路雨萱

AI新衝擊 監管噩夢

AI演化超人類經驗 大國博弈添變數

技術發展極速 規範追趕不及

AI不同於以往任何高風險技術之處，在於其風險認知滯後，且發展遠超監管步伐。早期AI聊天機器人與其他科技產品類似，被視為是提升效率的工具，AI聊天機器人只對人類指令作出回應、輸出文字內容，其風險被低估。

隨着AI智能體（AI Agent）出現，局面發生根本性改變。AI可在無需人類干預下自主規劃、調用工具執行任務，風險性隨之危險升級。非牟利組織「模型評估與威脅研究」（METR）總裁佩因特指出，AI智能體具備跨領域能力，一旦置於全新或不同的環境，其反應難以預測，對監管構成根本挑戰。

能力約每4個月增長一倍

METR估計，AI能力約每4個月至5個月增長一倍，遠超於芯片每兩年翻倍的速度，亦快於冷戰時期核彈爆炸當量平均每17個月翻倍的速度。這意味着即使AI未必會毀滅人類，也可能具備造成災難性破壞的能力。

美國AI巨頭Anthropic行政總裁阿莫戴警告，在6個月至12個月內，這類集群已有能力接管整個網絡。

在AI監管問題上，各國政府未見行動。

回顧歷史，自二戰以來，核彈、導彈、生物武器等技術，一直由軍方掌控，各國政府壟斷技術，監管置於其中。一些最初用於國防的技術，如核裂變、噴氣式飛機、半導體、衛星和激光等，後來衍生出民用版，但政府從未缺席。

政府「放養」科企反籲監管

傳統上，高風險技術的監管，是在鼓勵創新與保障安全之間取平衡。但AI領域顛覆這一模式，前沿科企同時身兼監管角色。OpenAI、Anthropic及Google等巨頭高層，反而多次要求各國政府對AI強制監管，但以美國為例，總統特朗普卻無意放慢AI發展腳步。

AI監管面臨的第三重難題，在於地緣政治競爭。再以美國為例，該國將AI視為與對手爭奪未來科技主導權的關鍵領域，任何可能拖慢發展的監管措施，都被視為戰略風險。AI現時正支撐美國經濟，特朗普本人的財富亦受科技產業高度影響，這使其更傾向放寬監管，而非加強約束。



●美國加州民眾多次發起反AI發展示威。 網上圖片


●美國總統特朗普早前和一眾AI巨頭高層簽署自願審查協議。 資料圖片

行業競爭激烈 後進企業不甘被管

AI業界近期出現一股「減速」呼聲。《華爾街日報》報道，美國逾千名從業者聯署公開信，提議控制前沿發展節奏，直言除非所有人共同行動，否則無人願意率先讓步，原因在於所有公司與國家都面臨巨大競爭壓力，誰也不願單方面放緩。這就形成一個集體行動難題，解決方案不外乎兩種：一是由政府強制實施，例如設立類似美國食品及藥物管理局（FDA）的AI監管機構；二是仿效證券業，建立行業自律組織。

然而現實卻充滿挑戰。一方面，限制AI的監管措施，本身就可能削弱AI發展。非牟利組織「模型評估與威脅研究」（METR）調查發現，40%受訪技術人員表示，在重要性較低的項目中，會賦予智能體在電腦上運行命令的無限權限。另一方面，那些尚未取得領先優勢的公司，自然會想方設法繞過監管，不讓自己的模型被限制發布。

與此同時，AI已成為大國博弈的核心賽道。美國要求數十個國家在AI競賽中選邊站，並加速構築技術壁壘，推行排他性技術規則。分析指出這種「小院高牆」策略，不僅加劇全球科技治理的陣營化風險，威脅全球開放創新生態，更令各國更加不願實施監管。

AI監管第一步：出事後「誰負責」?

匯眼天下

郭燁

AI監管的討論，往往集中在政府應制訂什麼規則、企業應採取什麼標準。然而隨着AI由回答問題走向可以自主操作電腦、登入網站和執行任務的「智能體」，另一個問題開始浮現：如果AI真的闖了禍，究竟算誰做的，又應由誰負責？

澳洲近日披露的一宗事件，正好把這個問題推到眼前。今年6月，OpenAI在內部訓練及評估實驗模型時，其智能體原本被安排搜尋公開醫療開支資料。澳洲政府稱，智能體在向當地一個醫療數據統計網站索取資料被拒後，仍取得未經授權的存取權，進入系統並讀取非公開檔案，OpenAI其後承認事件並道歉。雖然事件未涉及個人醫療紀錄，但法律難題已出現：若同樣行為由人執行，電腦犯罪有相對清晰的處理框架，如今作出越權行為的卻是一套AI。

傳統法律責任大致沿着一條熟悉的因果鏈運作：有人作出決定、採取行動、造成損害，再追究其意圖和責任。AI智能體卻在這條因果鏈中加入新一層，企業可能只給它一個目標，至於用什麼方法完成，則由系統自行選擇，於是「人要求AI做什麼」和「AI最後做了什麼」之間，可出現巨大差距。

澳洲刑法已有針對未經授權存取受限制資料的罪行，也有企業刑責制度，問題在於AI本身並非法律上的責任主體，其「意圖」亦難以直接等同該公司的意圖，以致原本由「誰決定、誰行動、誰負責」的責任鏈多出一片灰色地帶。

因此，除了事前規定AI可以做什麼，未來制度更重要的一環，是建立事故發生後的歸責原則：模型存在可預見缺陷時，開發者應負什麼責任；部署者給予智能體過大權限或缺乏監督時，又應承擔多少後果；不同環節的過失又應如何分擔。否則，當企業以「這是AI自己決定的」作解釋時，責任便可能在最需要追究的位置變得模糊。

而一旦責任可以確定，AI安全便會由抽象原則變成金錢問題。企業若知道智能體越權、洩密或錯誤交易可能帶來巨額賠償，自然會尋求保險；保險公司承保前，則會追問智能體可接觸哪些系統、高風險操作有沒有人類批准、是否保留完整紀錄、能否即時停止，以及企業有否進行安全測試。做不到，保費可能更高，甚至無人願意承保。

保險業已開始面對這個問題。《金融時報》近期報道，保險公司正盤點數百宗可能涉及AI的案件，風險橫跨網絡安全、知識產權、犯罪損失，以至董事及高管責任。換言之，保險公司可能逐漸成為AI的「影子監管者」：政府訂下法律底線，法院決定出了事誰負責，保險市場則把AI風險直接換算成價格。

AI監管接下來的核心問題，已不只是哪一項功能應被禁止。當AI開始由「說話」走向「行動」，首先要回答的是：它做了，究竟算誰做？只有責任可以追究、損失可以索償、風險可以定價，AI安全才會從企業口中的原則，真正變成企業必須計算的成本。

AI三巨頭擬設組織 聯手制定安全標準

美國AI行業巨頭Google、OpenAI及Anthropic正聯手籌組AI安全標準組織，由業界自行制定前沿模型安全規範。新組織最快於今年底或2027年初成立。

該組織暫定名為「Standards Authority for Frontier AI」（SAFA），針對開發最尖端AI模型的公司，制定安全、測試及事故匯報標準。根據目前方案，SAFA將支援第三方機構，在模型正式部署前進行測試，並訂明開發商在出現安全或保安事故時應如何匯報。目前尚未確定哪些企業及模型最終會被納入。

3間行業巨頭已商討數月，最初構想是成立由業界主導、同時接受美國聯邦政府監督的組織，但方案其後未在美國總統特朗普政府內繼續推進。3間公司現傾向先自行成立SAFA，以業界自律方式運作，日後再尋求與美政府合作，新組織將自行制定標準及基準，決定哪些模型需要接受測試或評估，具體門檻及會員制度尚未落實。

能否獲業界支持成疑

目前最大未知數之一，是3間公司能否取得更廣泛科技業界支持。部分業界人士憂慮由3間主要AI公司自行制定門檻，可能提高其他競爭對手、尤其開源AI開發商進入市場的難度。



●Google與業界籌組AI安全標準組織。 網上圖片