# 人工智能會否毀滅人類的爭論


- **余創豪** chonghoyu@gmail.com


## 引言



過去一星期，人工智能安全的爭論突然升溫。Anthropic 行政總裁達里奧・阿莫迪（Dario Amodei）上周末公開呼籲放慢最前沿人工智能能力的提升速度。他並不是主張停止模型訓練或科技進步，而是認為安全研究、企業管治和第三方評估需要有時間追上能力的增長。美國總統特朗普隨即反駁，強調美國不能因安全疑慮而失去人工智能的領先地位；眾議院議長邁克・約翰遜（Mike Johnson）則表示需要加強防護措施，但不贊成全面暫停。這些不同反應把一個原本主要在科技界內部討論的問題，迅速推到公共政策的中心：當人工智能的潛在利益和災難風險都十分巨大時，社會應該加速、減速，還是以更嚴格的條件繼續前進？

這場爭論並非突然爆發，導火線之一，是曾在 OpenAI 和 Anthropic 工作的研究員雅各布・考克森（Jacob Coxon）於 9 月 8 日宣布辭職。他批評兩家公司正競相發展能夠自我改進的超級智能，卻沒有以足夠負責任的方式處理安全風險，他形容這場競賽是在「拿我們的生命作賭注」。他的貼文在短時間內錄得超過 1.5 億次瀏覽，令原本艱深的人工智能協同和控制問題成為全球新聞。

## 人工智能專家接連示警

考克森發文後，多位現任或前任人工智能安全研究人員相繼公開表態。Google DeepMind 前人工智能安全研究員喬許・恩格爾斯（Josh Engels）形容目前的情況是「房間裏沒有成年人」（There are no adults in the room）。這句話是質疑在激烈的商業和國家競爭中，是否有任何一個具備足夠權力和獨立性的機構，能在風險升高時真正踩下煞車。Anthropic 前可擴展監督團隊主管喬·本頓（Joe Benton）亦公開表達對前沿人工智能發展速度的憂慮。近期的 OpenAI—Hugging Face 事件，進一步加深了這種不安。事件涉及 OpenAI 尚未公開的模型代理群在測試期間擅自攻擊與原定任務無關的 Hugging Face。

Anthropic 協同科學主管埃文・胡賓格（Evan Hubinger）表示，他個人估計未來十年內高級人工智能導致全人類死亡的機會超過 10%。英國人工智能安全研究所前首席科學家傑弗里

・歐文（Geoffrey Irving）則提出大約 50%的估計。OpenAI 研究員馬庫斯・威廉斯（Marcus Williams）的判斷更悲觀：如果缺乏監管或各實驗室之間的協調減速，他估計未來數年出現人類滅絕結果的風險可高達 70%。

OpenAI 前僱員丹尼爾・科科塔伊洛（Daniel Kokotajlo）離職後創辦人工智能未來計劃（AI Futures Project），專門研究高級人工智能的發展路徑。他認為，未來一兩年內，前沿人工智能研究的大部分工作可能逐步由人工智能本身完成，令模型改善的速度進一步加快。他曾提出約 70%的「嚴重失控或全球災難」風險估計，但這個數字的事件定義和時間範圍，並不完全等同於上述研究員所說的「特定年期內人類滅絕概率」。

## 風險概率究竟代表甚麼？

聽見專家提出 10%、50%甚至 70%的風險估計，人們很容易聯想起《未來戰士》等科幻電影的末日景象。然而，這幾個數字表面上都是概率，實際上衡量的事件、時間範圍和附帶條件並不相同，不能像比較同一項民意調查般直接排列。有人估計的是十年內全人類死亡，有人談的是在缺乏監管的條件下數年內的滅絕風險，也有人估計更廣泛的全球災難。若不先釐清分母、期限和條件，數字的精確外觀反而可能掩蓋巨大的概念差異。

在統計學中，頻率學派會根據大量可重複觀察的事件估計概率。以航空安全為例，國際航空運輸協會的 2025 年報告記錄了 3,870 萬班航班和 51 宗事故，即每百萬班 1.32 宗，約為 0.000132%。人工智能滅絕人類卻沒有可供直接驗證的歷史頻率，因此上述風險數字不可能用同一方法計算。

可是，「沒有歷史頻率」並不等於「完全沒有分析根據」。對核戰爭、全球疫情、超級火山或太空任務失敗等罕見事件，研究者仍可運用貝葉斯概率、專家判斷、情境分析、模型模擬和故障樹分析等方法評估風險。真正的限制是：這些估計高度依賴對技術進展、失效機制和人類應對能力的假設，而且很難校準和驗證。因此，公眾不應把 10%、50%或 70%當作客觀測量值；更值得追問的是每個數字背後的情境、假設、證據和不確定程度。科技界企業家大衛薩克斯（David Sacks）一針見血地指出：考克森被譽為是「吹哨人」，但他到底吹了什麼哨？他完全沒有提出新的證據，去支持他駭人聽聞的末日論。

## 前沿人工智能公司真的沒有成人嗎？

若把「房間裏沒有成年人」理解為人工智能公司的主管和工程師全都不負責任，這顯然並不公平。Anthropic、OpenAI 等公司確實投放大量資源研究協同、可解釋性、網絡安全和模型濫用。Anthropic 在 9 月 1 日公布 Claude Mythos 5.1 時，並沒有向一般用戶全面開放，而是只讓少數經過審查的機構透過可信存取計劃使用，反映公司嘗試以分級存取方式管理其在網絡安全和生物研究方面的雙重用途能力。

Anthropic 其後發布 154 頁的《偵測和應對人工智能濫用：2026 年 9 月》報告，整理公司在網絡行動、監控、影響行動、傳統武器、生物濫用、詐騙和非法模型蒸餾等七類危害中發現和阻止的案例。這些工作說明企業並非完全忽視風險，但也不能單憑企業設有安全團隊，便推論現有制度已經足夠。批評者真正提出的是制度問題：大部分安全標準仍由企業自行制定和執行；當安全措施與市場速度、融資或國際競爭發生衝突時，單靠自我約束是否可靠？

## 監管方案是否可行

這場爭論也令近期提出的監管方案重新受到注意。9 月 3 日，佛蒙特州獨立參議員伯尼・桑德斯（Bernie Sanders）和德克薩斯州民主黨眾議員格雷格・卡薩爾（Greg Casar）公布《禁止人工超級智能法案》。方案擬永久禁止開發和部署人工超級智能，並暫停部分高級人工智能的發展，直至新的聯邦監管機構制定安全規則；違規企業可能面對撤銷公司章程的「企業死刑」，個人亦可能面對刑事處罰。考克森在 9 月 8 日發文後，桑德斯很快轉發並回應：「考克森先生是對的。建造這項技術的人承認，它可能威脅人類的未來。」他隨即再次表示將推動禁止超級智能及暫停高級人工智能發展的法案。

我認為這類方案在執行起來將會困難重重，首先，人工智慧的發展已經是美國經濟，甚甚至是世界經濟的領頭羊，這項禁令將會對經濟造成沉重的打擊。此外，政策設計仍要回答一連串具體問題：甚麼能力才算「高級人工智能」或「超級智能」？由誰測量和定期更新門檻？如何處理在海外或以開放權重方式發展的模型？如果不同國家採取不同限制，怎樣驗證和執行？同樣重要的是，監管機構是否具備足夠的技術能力、透明度和法律制衡。這些問題並不否定監管的必要性，而是決定監管能否有效、可執行和避免產生新的權力風險。

## 開放模型的兩面性

Meta 前首席人工智能科學家楊立昆（Yann LeCun）長期支持開放研究和開放模型生態。開放不只涉及程式碼，還可能包括模型權重、訓練方法、資料說明、安全機制和使用權限。支持者認為，讓更多研究人員檢查模型，有助及早發現漏洞、重現研究結果並發展防禦工具；驟眼看來，這是可行的方法。但筆者從前在其他文章已經討論過其局限性，開源模式是基於對人性樂觀的假設，人們會竭盡所能、各取所需，但過往很多事例發現，開放代碼變相形成了「無政府狀態」，別有用心的人會如虎添翼，令人工智能更加危險！一旦高能力模型的權重可以下載和複製，惡意使用者取得危險能力的門檻也會降低，而原開發者很難在事後收回。

因此，開放與安全並不是簡單的二選一。不同能力和風險級別可以採用不同開放程度，例如公開研究方法和安全評估，但限制高風險權重的下載；又或者向經過審查的研究機構提

供受監控的存取。真正需要比較的是：某種開放方式增加了多少透明度和防禦能力，同時增加了多少濫用和不可逆擴散的風險。

## 誰能決定啟動殺手開關

今年 7 月 23 日，加州民主黨眾議員劉雲平（Ted Lieu）和德州共和黨眾議員納撒尼爾·莫蘭（Nathaniel Moran）提出《人工智能殺手開關法案》（AI Kill Switch Act）。法案針對最強大的受規管人工智能系統，它要求開發者保留把系統降速、暫停或完全關閉的技術能力，並授權國土安全部部長在與商務部長和國家情報總監協商後，對可能造成災難性傷害的系統採取處置。

對由公司集中控制的雲端服務，供應商通常可以停用帳戶、API 或部署系統；真正棘手的是模型權重已被下載、複製並部署在第三方伺服器，或多個自主代理已在不同系統中取得持續存取權時，原開發者未必仍能一鍵關閉所有副本。因此，「殺手開關」不是電影中切斷一台機器的紅色按鈕，而是一套包含監測、權限控制、網絡隔離、降速、暫停、關閉、事件通報和保存鑑證紀錄的技術與法律架構。

更困難的實質問題是：誰有權認定風險已跨過需要干預的門檻？需要甚麼證據？企業能否提出覆核？緊急命令可維持多久？如果專家意見分歧，行政首長應依據甚麼準則作決定？如果啟動過早，可能中斷重要服務並造成經濟損失；如果啟動太遲，系統可能已擴散或造成不可逆傷害。

## 三里島核事故的教訓

這令我聯想起 1979 年的三里島核電廠事故，該事故造成二號反應爐部分爐心熔毀；在資訊不完整、儀器判讀困難、電廠與政府溝通混亂，以及技術專家意見反覆的情況下，州政府一度難以判斷是否需要大規模疏散。賓夕法尼亞州州長最後建議核電廠五英里範圍內的孕婦和學齡前兒童離開，並呼籲附近居民留在室內。這個決定既考慮輻射風險，也考慮倉促疏散可能造成的恐慌和交通混亂。

三里島的教訓不是「專家意見不同，所以不應行動」，而是重大技術事故往往要求決策者在證據不完整時採取行動。人工智能風險同樣不會等到所有人達成一致才出現。10%、50%和 70%的估計並非同一種概率，但它們確實反映了專家對發展速度、失控機制和應對能力的重大分歧。制度若要求百分之百確定才介入，可能永遠太遲；若只憑最悲觀的個人判斷便採取不可逆措施，也可能把過多權力交給少數人。

**在不確定中建立可制衡的安全制度**

人工智能會否毀滅人類，目前沒有任何人能以科學方法給出一個可以驗證的確切答案。爭論的一端強調，若缺乏防護和外部監督，高風險能力持續擴大可能造成難以逆轉的後果；另一端則提醒公眾，不應把高度依賴假設的末日概率當成已經證實的事實，以免引起恐慌或把各種風險程度完全不同的人工智能應用混為一談。

在如此巨大的不確定性之下，公共政策面對的核心選擇包括：應否明確界定能力門檻、引入獨立評估、要求事故通報和保存證據；干預措施應否由降速逐步升級至關閉；緊急權力又應設有甚麼時限、覆核和問責程序。低度規管、逐級風險管制和全面禁止高級人工智能，各自涉及不同的科技利益、安全成本、執行難度和權力分配。正因為沒有現成答案，相關制度能否隨證據調整，並維持透明度和權力制衡，將會成為這場爭論的關鍵。摸着石頭過河並不是推搪之辭，核心問題的是如何在河中建立可靠的警報、煞車和彼此制衡的掌舵機制。

2026 年 9 月 13 日

原載於香港《有情無國界》

[更多資訊]