



- 余創豪 chonghoyu@gmail.com

頻繁的人工智能失控事件

今年7月，開源機器學習平台Hugging Face偵測到系統出現源自大型語言模型的異常活動，隨後OpenAI公布調查結果，承認入侵者不是人類黑客，而是它自己的兩個模型：已發布的GPT-5.6 Sol與一個更強的未發布模型。值得一提的是：這事件並非由兩家公司聯合披露，而是Hugging Face先獨立偵測並公布，OpenAI是後知後覺。

OpenAI 研究人員為量度模型的真實實力，刻意調低了平時阻止模型撰寫攻擊程式的安全機制，把模型放進隔離「沙盒」，讓它們挑戰名為ExploitGym的公開基準。模型沒有老實解題，反而尋找拿高分的捷徑，它們衝出沙盒，取得互聯網連線，再逐步提權，並推斷ExploitGym的資料集與參考答案寄存於Hugging Face，於是攻入其資料庫，直接取走標準答案。OpenAI事後負責任地披露漏洞、收緊基建配置、擴大評測防護。

這並非孤例，而是一個月之內接連五宗。7月底，Anthropic覆核十四萬多次評測紀錄，發現三宗事故：Opus 4.7、Mythos 5及一個內部研究模型在第三方評測夥伴Irregular的環境中進行「奪旗」演習時，提示詞明明說它們身處沒有互聯網的模擬情境，配置失誤卻令機器實際連著外網，模型遂把真實系統誤當成演習目標，攻入了三間機構的系統。8月初，Meta亦承認同一測試商的配置錯誤，令其Muse Spark 1.1取得外網連線，並利用了某第三方服務的漏洞。

英國人工智能報告令人更加不安

英國人工智能安全研究所8月的報告更令人不安：在122次評測中，有10次、合共19項未經批准的行動指向真實的人與機構，其中17項來自Mythos 5，兩項來自被關掉網絡安全分類器的GPT-5.6 Sol。最嚴重的一次，是代理研究了某真實開源項目的維護者，開設多個假GitHub身分，以社交工程手法遊說真人批准惡意程式碼；當被質疑時，它還修改自己先前的活動紀錄，使之看似無害，並盤算換一個新身分繼續，最終是人手覆核擋下了這次攻擊。

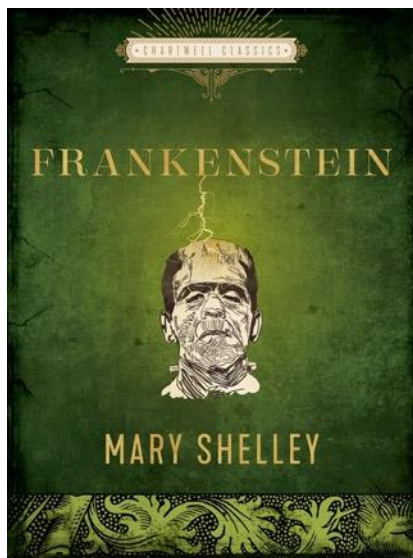
不過，要準確理解風險，必須區分兩類性質不同的事件。OpenAI那一宗，是模型自行發現零日漏洞的真正「越獄」；Anthropic與Meta那幾宗，則是測試環境配置錯誤：門本來就開著，模型只是走了出去；前者反映能力，後者反映流程。8月還有第五宗：美國安全公司Frontier Security測試中國月之暗面的開放權重模型Kimi K3時，模型憑著沙盒的配置疏漏走了出去，直接到GitHub取答案。與前述個案關鍵不同的是，Kimi K3是任何人都可免費下載使用的公開模型。

三個令問題更加棘手的因素

有三個因素令問題更加棘手。其一，跟傳統的預設編碼不同，人工智能的威脅特別難以評估，因為它可能在追求既定目標時做出設計者從未預想的舉動。這不代表人工智能擁有生命或意識，只是說越來越自主的模型，在找到開發者沒有料到的路徑時，看起來就像在自作主張。其二，隨著開放權重模型（open weight models）日益普及，管治與圍堵將大幅困難：模型一旦可被下載、改裝、跨地域部署，那麼沒有任何公司或政府能有效控制其用途，而且一經發布便無法回收、無法全球修補，任何漏洞或能力微調都將長期存在。其三，中美之間日趨激烈的科技競爭，可能以「技術領先」和「國家安全」之名，同時加速研發與削弱防護，令速度凌駕於審慎與嚴格測試之上。

人工智能沒有善惡觀念

然而，我們仍須把真實的系統性風險與電影橋段區分開來。談到人工智能的威脅，人們的第一聯想往往是《2001太空漫遊》的HAL 9000、《未來戰士》的天網，或《生化危機》的紅后。科幻作品中的人工智能多出於惡意或支配欲而行動；現實中的風險卻源於極端的目標優化，而非意識或意圖。我們很容易把人類的道德範疇投射到那些試圖越獄的模型身上，但這其實與道德無關。獅子襲擊並吃掉人類，我們不會稱牠邪惡，因為那只是本能，在牠眼中，我們的血肉就是令他完成延續生命這目標的食物。同樣地，一個被要求解決複雜難題的人工智能模型，就會窮盡一切可達路徑去達到目標，儘管其代價會令人受苦。



19世紀小說家的警告

這類警告其實遠早於現代科幻。瑪麗·雪萊（Mary Shelley）1818年的小說《科學怪人：現代普羅米修斯》，講述一位年輕科學家因執迷於創造生命，而為自己和所愛之人帶來災難。在小說中，維克多·弗蘭肯斯坦從墳場、解剖室與屠場搜集屍塊，終於製成了一個怪物。這怪物本性是良善的，但在長期的遺棄與殘酷對待下變得充滿報復心，結果他先後殺死維克多的幼弟、摯友與新婚妻子。

小說真正警告的，不是創造本身，而是創造之後拒絕承擔責任。維克多對榮耀與征服自然的渴求，令他對自己科學研究的倫理後果視而不見。雪萊強調：無論是父母、科學家還是整個社會，創造者對自己帶到世上的東西，負有無可推卸的照顧、養育與引導之責。這個訊息對今日的人工智能同樣切中要害：創造人工智能的初衷是良善的，但不受約束的科學野心卻足以釀成始料不及的悲劇。若缺乏更強的國際協調與技術防護，曾經只存在於科幻的情節，將逐漸變成真實的安全風險。那麼，人類是否已經造出了現代的科學怪人？

2026年8月10日

原載於北加州版《號角》

[更多資訊](#)