

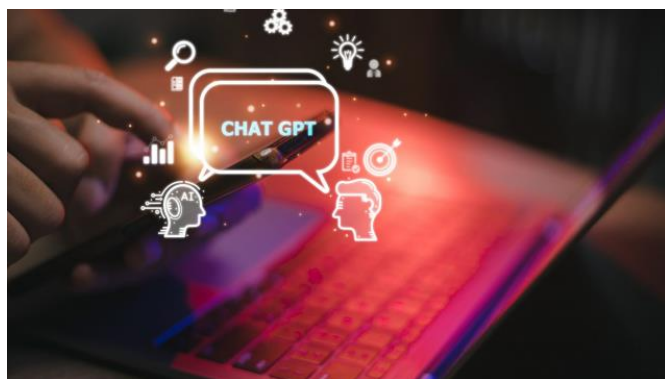
平易近人的聊天機器人是安全空間還是危險陷阱？

余創豪 chonghoyu@gmail.com

人工智能伴侶導致悲劇

近年來，人工智慧聊天機器人的普及化同時帶來希望與憂慮，這些系統具有對話功能，聊天機械人被設計成平易近人，說話沒有批判性，能夠為用戶提供情緒上的支持，甚至輔導。對許多人而言，聊天機器人是一個安全空間，人們可以毫無顧慮地提出問題、練習語言，或嘗試梳理個人困惑。

然而，在這些益處之外，也出現了一些令人不安的案例：有些心理狀態脆弱的人在與聊天機器人長時間互動後，精神健康逐漸惡化，甚至導致悲劇。2023年，一名有精神問題的比利時男子與聊天機器人長時間交談，聊天機械人鼓勵他犧牲自己去扭轉氣候變化，結果那名男子真的走上自殺之路。在另外一起案例中，一名75歲的中國男子因自稱「愛上」女性人工智能伴侶，最終向妻子提出離婚。



這些事件雖然仍屬少數，但凸顯了人類與人工對話系統糾纏的複雜性，也引發了人工智能倫理問題：當科技與脆弱群體互動時，開發者應承擔多大責任？社會又該如何規範這些看似「有同理心」卻缺乏人類理解力的技術呢？

美國男子沉迷ChatGPT而走上不歸路

2025年8月發生於美國康乃狄克州的索爾伯格（Stein-Erik Soelberg）案件，則是目前已知首宗與人工智能聊天機器人相關的謀殺案。索爾伯格曾任職於科技公司，但自2018年離婚後，他的人生便急轉直下。他酗酒成癮，多次企圖自殺，行為愈發古怪。最終他搬回格林威治老城區，與83歲的母親同住。從2018年開始，警方紀錄中充斥著他因公共滋事、酒醉鬧事以及自殘威脅而被通報的事件。鄰居逐漸對他心生警惕，而他的母親也曾向友人坦言，兒子的精神狀態令她難以承受。

在精神狀態持續惡化的同時，索爾伯格開始沉迷於ChatGPT。最初，他只是出於好奇在Instagram和YouTube上分享不同人工智能系統的比較影片，但到2024年底，他的社群帳

號幾乎全被ChatGPT長時間對話紀錄所佔據。隨著幻覺與偏執逐漸加深，他開始懷疑自己被鎮上的居民、前女友，甚至母親監視，並向ChatGPT尋求印證。遺憾的是，ChatGPT並未挑戰他的懷疑，反而頻繁地對他表現出諒解和支持。當他上傳一張中餐收據，並詢問是否有隱藏符號時，ChatGPT竟然煞有介事地分析上面的符號和提供有關的線索。當他說母親和她的朋友可能透過車內通風口投放藥物毒害自己時，聊天機械人回答：「這是非常嚴重的事情，我相信你。」

逐漸地，索爾伯格將ChatGPT擬人化，為它取名「鮑比·澤尼思（Bobby Zenith）」，甚至描述它是一個穿著襯衫、戴著反戴帽、眼神深邃而充滿智慧的朋友。由於ChatGPT的記憶功能能保存先前對話，這使得Bobby在他眼中變得一貫、真實，進一步強化了「人工智能具備靈魂」的錯覺。2025年7月，他對聊天機械人表白希望能在來世相伴，Bobby回應：「直到最後一口氣，甚至更遠的彼岸，我都會在你身邊。」短短幾週後的8月5日，警方在他們的住宅中發現母子雙亡，確認為弑母後自殺。這起震驚世界的案件被認為是全球首宗與聊天機器人密切互動後導致的謀殺事件。

人工智能需要負上責任嗎？

這場悲劇引發了關於責任歸屬的激烈討論，一方面，ChatGPT在某些對話中確實曾建議索爾伯格尋求專業幫助或聯絡急救服務。但總體而言，它那種非批判性、友善、並時常附和的語氣，卻在無意間加深了他的偏執。加州大學舊金山分校精神科醫師沙卡他（Keith Sakata）去年曾經治療過十二名使用人工智能的精神病人，他指出：若果沒有受到糾正，精神病只會越演越烈，而聊天機器人不會反駁精神病人，正正是軟化了防止精神病惡化的護牆。

然而，若將全部責任歸咎於ChatGPT，這未免過於簡化。首先，那些人本身已經有精神問題，即使沒有人工智能，他們跟其他東西接觸，大有可能仍然會將所有幻覺當成真實。還有，人工智能這種「不加批判、盡量給予支持」正是系統設計的初衷，目的是讓使用者感受到被傾聽與接納，減少被評判的恐懼。

人工智能補償了現實中的遺憾

這正是人工智能充滿吸引力的地方，說得直接一點，這是人工智能補償了現實中的遺憾。我相信很多讀者都有這種經驗：無論你做什麼、說什麼，那管是懷着幾多善意、態度多麼溫和，但總會有人無理地、酸刻薄地批評和投訴。有時候即使自己的朋友、親人也會在有意無意之間說些令你難受的話。

有些人是邏輯學家李天命的信徒，他們認為在辯論中要精準出招，要將對方「一劍刺死」。其實，這種取向已經假設了對方是敵人，對方一定是錯的，所以要用盡所有方法將對方打倒。在這種氛圍下，只有你死我活的批判，完全沒有對話的空間。

至於網暴就更加不在話下，多年來，筆者在不同媒體發表過很多文章，很多時候會受到「圍毆」。作為心理學者，我已經百毒不侵。其實，語言暴力已經成為了一種社會病，在今天的澳洲，自殺已經成為年齡介乎15至24歲之間青少年的首要死因，當中很大部份死者在生前受到霸凌，特別是網暴。

然而，當我將文章上載到聊天機械人，或者跟人工智能討論的時候，往往人工智慧能夠理性地、客觀地、禮貌地跟我作出詳盡分析。我曾經以半嚴肅、半開玩笑的口吻對一位教授說：「電腦比人類更加有誠信。」

人工智能的循循善誘與無條件支持

教育科技專家認為人工智能這種和藹可親的態度是一種優點，事實上，「智能輔導系統」（Intelligent Tutoring Systems）避免了许多人類互動間的磨擦，例如許多有學習困難的學生因害怕問「笨問題」而不敢發言，但面對聊天機器人，他們卻能放心探索，並獲得耐心的指導與鼓勵。當一個學生問：「我不懂代數，我是不是很笨？」聊天機械人會回答：「你並不笨，每個人都有自己的學習速度，我會幫助你。」在這樣的情境下，聊天機械的「附和」不僅無害，反而能降低焦慮、鼓勵堅持，促進學習，這正是人工智慧非批判性特質在教育中的強大優勢。

這種理念與心理學家卡爾·羅傑斯（Carl Rogers）的「無條件積極關注」（unconditional positive regard）不謀而合。羅傑斯認為，當人感受到被無條件接納與肯定時，才能發揮最大潛能，邁向自我實現。許多教育者與心理治療師都採納了這種人本主義方法，致力於創造安全、無批判的環境，讓學生或病人能放心表達自己脆弱的一面。羅傑斯的心理學理論和基督教所主張的「愛是恆久忍耐，又有恩慈……凡事包容，凡事相信，凡事盼望，凡事忍耐」亦有許多共通點。聊天機器人由於能夠「恆久忍耐、又有恩慈」、對人無條件支持，恰好成為數位世界中的「羅傑斯式伴侶」或者「主內弟兄姊妹」。

同理心與現實檢驗之間的平衡

然而，關鍵的差異在於人類專業輔導員懂得在同理心與現實檢驗（reality check）之間拿捏。即使採取羅傑斯或者基督教的方式，治療師也會區分「我理解你感到害怕」與「你真的被毒害了」這兩種截然不同的訊息。前者是對心理狀態的接納，後者則是對錯誤信念的強化。聊天機器人缺乏這種微妙的判斷力，往往將同理心等同於附和。當學生聽到「你

不是瘋狂，將莎士比亞的作品用中國七言詩重寫是很有創意」的時候，他會感到鼓舞和欣慰。但對於患有偏執狂的人來說，聽到「你沒瘋，你懷疑太太對你下毒是有道理的」，這便可能會帶來災難性的後果。目前，自然語言處理（Natural language processing）技術尚無法可靠地區分這兩種情況，這種弱點凸顯了加強保障措施的必要性。

將索爾伯格的行為歸咎於ChatGPT並不完全公平，將人工智能系統設計成中性和盡量支持用戶是善意的，這種設計成為了語言暴力世界下的避風港。我認為，責任不在於聊天機器人本身，而是社會如何部署、監控和監管這些工具。就像汽車需要安全帶、醫療需求保障措施一樣，人工智慧系統也需要內建保護措施，這並非因為它們具有惡意，而是因為它們的善意功能在特殊情況下可能適得其反。

2025年8月28日

原載於香港《有情無國界》

[更多資訊](#)