- 余創豪 chonghoyu@gmail.com

近來不少生成式人工智能（Generative AI）公司都因為法律訴訟而疲於奔命。今年七月，美國喜劇演員莎拉・西爾弗曼（Sarah Silverman）和兩位作者在三藩市地方法院對 OpenAI 和 Meta 提起訴訟，指控這兩家公司為了訓練其人工智慧程式，複製了他們受到版權法保護的作品。

今年九月底，在作家協會的支持下，十七位美國作者向紐約聯邦法院起訴 OpenAI 侵犯版權，因為在未經作者同意下，OpenAI 採用他們的作品去訓練其人工智能系統，並且為用戶提供答案，但他們分文也收不到。那些作者指責 ChatGPT 這是「大規模的系統性盜竊」。

無獨有偶，幾乎在同一時間，澳洲作家亦提出同樣的指控，澳洲作家李察・弗拉納根（Richard Flanagan ）發現他的十部小說被 Meta 和 Bloomberg 等公司用來訓練生成式人工智慧，他稱之為「史上最大的版權盜竊行為」。 他說：「我感覺自己的靈魂彷彿被剝開了，但我無力阻止。」澳洲出版商協會表示，多達一萬八千本澳洲的書籍可能被生成式人工智慧侵權。

我並不是律師或者法律專家，在這裏我只是提出一點外行人的淺見，這篇文章的重點是通過這些事件，去討論生成式人工智能是不是一部大型的抄襲機器。

以下是一本書典型的版權聲明：「本書的任何部分不得以任何形式或任何方式複製、傳播，包括影印、錄音或透過任何資訊儲存和檢索系統，但在書評中使用簡短引文除外。」其關鍵詞是「複製、傳播」，如果我在圖書館把一本書影印了幾十頁，或者在圖書

館的網站下載了一本電子書的幾個章節，但我並沒有將資料轉發給任何人，而是寫論文時用來參考，這當然並不算侵權。若果我閱讀了幾十本書之後寫成論文，然後在期刊上發表文章，這也不算侵權，因為我並不是一字不改地抄襲。簡單地說，只是複製而沒有傳播，或者傳播但並不是複製，這都沒有問題。

請各位讀者想像以下的情況：我有幸獲得了美國全國科學基金會的研究經費，於是我聘請了三十名研究助理，首先，我向他們派發了幾本書和幾十篇論文，從而訓練他們的研究方法，跟着吩咐他們在圖書館和互聯網搜尋資料，然後將資料存入電腦，最後綜合成一份報告，在這過程中我們並沒有徵求作者的同意。令一個類似的情景是：某網紅同樣聘請了幾十名研究助理，他們亦是上窮碧落下黃泉地蒐集資料，那位網紅根據助手整理出來的資料拍攝 YouTube 影片，他們也沒有得到作者的同意。我很難想像這裏出現了侵權問題。

現在，如果我們將「研究助理」改為「生成式人工智慧」，那麼兩者的分別在哪裏呢？ 可能會有人反駁說：「撰寫論文需要提供注釋和參考書目，但生成式人工智慧沒有這樣做。」 雖然一些 YouTube 影片、報章、雜誌等傳媒偶爾會提及資料來源，但在大部份情況下並不會提供注釋和參考書目。

聊天機械人是賦予人工智能的研究助理，它們並不是搬字過紙，而是具有分析和綜合的能力，往往生成的內容和原來的資料有很大的不同，從這個角度來看，這些內容是具有「轉化的價值」（Transformative value）。 然而，有些人卻誤會了生成式人工智慧是抄襲的機器，例如阿蘇撒太平洋大學（Azusa Pacific University）哲學系教授米赫雷圖・古塔（Mihretu Guta） 在今年六月一個學術會議上指出： ChatGPT 只是翻抄現存的知識，採用 ChatGPT 只會令學問停滯不前；今年八月，紐約城市大學物理系教授加來道雄在接受有線新聞網絡採訪時說：ChatGPT 只是一台「 美化了的錄音機」（Glorified tape recorder）。其實，人工智能是可以通過「模式識別」（pattern recognition）而學習、甚至創造新知識，不過，我不想在這篇雜誌文章中說得太複雜。

退一步說，即使人工智能並不能好像人類般學習和創造新知識，但至少它可以幫助發明家縮短研發的時間。創新並不是在真空中爆發出來的，而是需要建基於過去汗牛充棟的資料， 那麼，人類面臨的困境便正如莊子所說：「吾生也有涯，而知也無涯。」

在未有人工智慧之前，學者只能夠依靠自己的博聞強記，舉例說，1986 年芝加哥大學圖書館學院長斯旺森（Don Swanson）仔細審查醫學文獻，嘗試去找出看似無關的事件之間的關係，在浩如煙海的論文中，他察覺到食用魚油、血液粘度降低、雷諾氏病可能有關，最終他的假設得到實驗研究的證實。換言之，翻查舊文獻是可以發現新知識的！人們稱這種在文獻中發掘概念關係的方法為「斯旺森過程」（Swanson process）。起初學者只能夠倚賴人去閱讀文獻和識別模式，後來人們採用了人工智能的「自然語言處理」（Natural

language processing），這種新方法名為「文字開礦」（Text mining），這是電腦版的「斯旺森過程」，不同之處是：傳統的斯旺森過程需要人力，但文字開礦則借助人工智慧。有趣的是，文字開礦已經至少有十多年歷史，但從來未發生過法律訴訟。

聊天機械人是文字開礦的延伸，它的威力更大，而且進入了主流，這引起了作家的反彈是不足為奇的。ChatGPT、Bard、Claude 等生成式人工智慧可以將斯旺森過程推展到極致， 從前研究助理可能需要花幾十個小時才獲取到需要的資料，現在只需要幾秒鐘，類似斯旺森的思想家便可以更加有效率地發現更多新知識！

不過，我理解到那些作家的感受，我猜想這些法律訴訟將會改變了現存的版權法和人工智能公司蒐集資料的界限。坦白說，我並不介意自己的作品被收錄在生成式人工智能的資料庫中，如果我有份貢獻於創造出新知識，我會感到萬二分榮幸！

2023 年 10 月 1 日

原載於澳洲《同路人》雜誌

[更多資訊]