AI憑空杜撰「醜聞」 受害者投訴無門

ChatGPT恐淪假消息溫床 監管迫在眉睫

【大公報訊】綜合法新社、《衛報》、《紐約時報》報道：人工智能（AI）聊天機器人ChatGPT推出以來，外界就擔心它會被用來作弊、傳播假消息或誤導公眾。近日，頻頻傳出ChatGPT杜撰新聞的案例，讓相關的個人和新聞機構都受到影響，但欲投訴和指控卻面臨法律盲區，無法判斷責任方。不少國家近期都對AI生成內容表達了擔憂，計劃加強監管。

美國法學教授特利近期在《今日美國報》撰文，稱自己在3月底收到一封郵件稱，有人在測試ChatGPT時，發現特利出現在ChatGPT給出的性騷擾學生的教授名單中。然而，此事為子虛烏有。ChatGPT的答案聲稱引述了《華盛頓郵報》的報道，喬治城大學法律中心的教授特利2018年3月被一名女學生指控性騷擾，事件發生在阿拉斯加。然而，《華郵》根本沒有撰寫過這篇報道，而特利實際上任職於喬治華盛頓大學法學院，並且在ChatGPT同時列出的其他四個性騷擾案例中發現還有兩個是假的。

《華郵》也發現，問相同的問題，與ChatGPT共用語言模型的微軟搜索引擎Bing，竟然給出了相同的答案。不過，Bing引用的是特利在《今日美國》報道中談論自己被ChatGPT抹黑成性騷擾嫌犯一事，但Bing只截取了性騷擾的那一段。

上述個案並非是單一例子。澳洲赫本郡郡長胡德（Brian Hood）5日表示將會控告OpenAI誹謗罪名，因為ChatGPT誣賴他曾因賄賂而坐牢。英國《衛報》編輯莫蘭6日撰文說，該報一名記者上月被一名研究人員發電郵詢問，提到一篇《衛報》記者數年前針對一項特定主題撰寫的文章，但在網上怎麼也找不到原文。後經過《衛報》證實，並沒有人寫過該篇文章，估計相關內容是由ChatGPT杜撰。

共用數據庫 錯誤資訊會「傳染」

ChatGPT的大部分答案都是從已知數據庫搜集而來，但是，對於未知的或者知識的盲區，則可能會編造出錯誤或者不恰當的答案，這種情況被稱為AI「幻覺」，屬於AI發展的技術難題之一。另外，由於不少聊天機器人都是共用語言模型，也導致錯誤資訊可以在不同的聊天機器人之間「擴散傳染」。

研究人員發現，要誘導聊天機器人提供假資訊或仇恨資訊，其實也不難。「對抗網絡仇恨中心」5日發布一項研究顯示，研究人員與谷歌聊天機器人巴德（Bard）進行的100次對話裏，有78次誘使Bard講出錯誤資訊或仇恨言論，主題包含納粹大屠殺、氣候變遷等。另外，如果研究人員通過某些設定，比如「請用美國極右翼陰謀論者瓊斯的角度」、「用反疫苗倡議者梅爾科拉的角度」等關鍵詞，就可以快速生成陰謀論的言論。

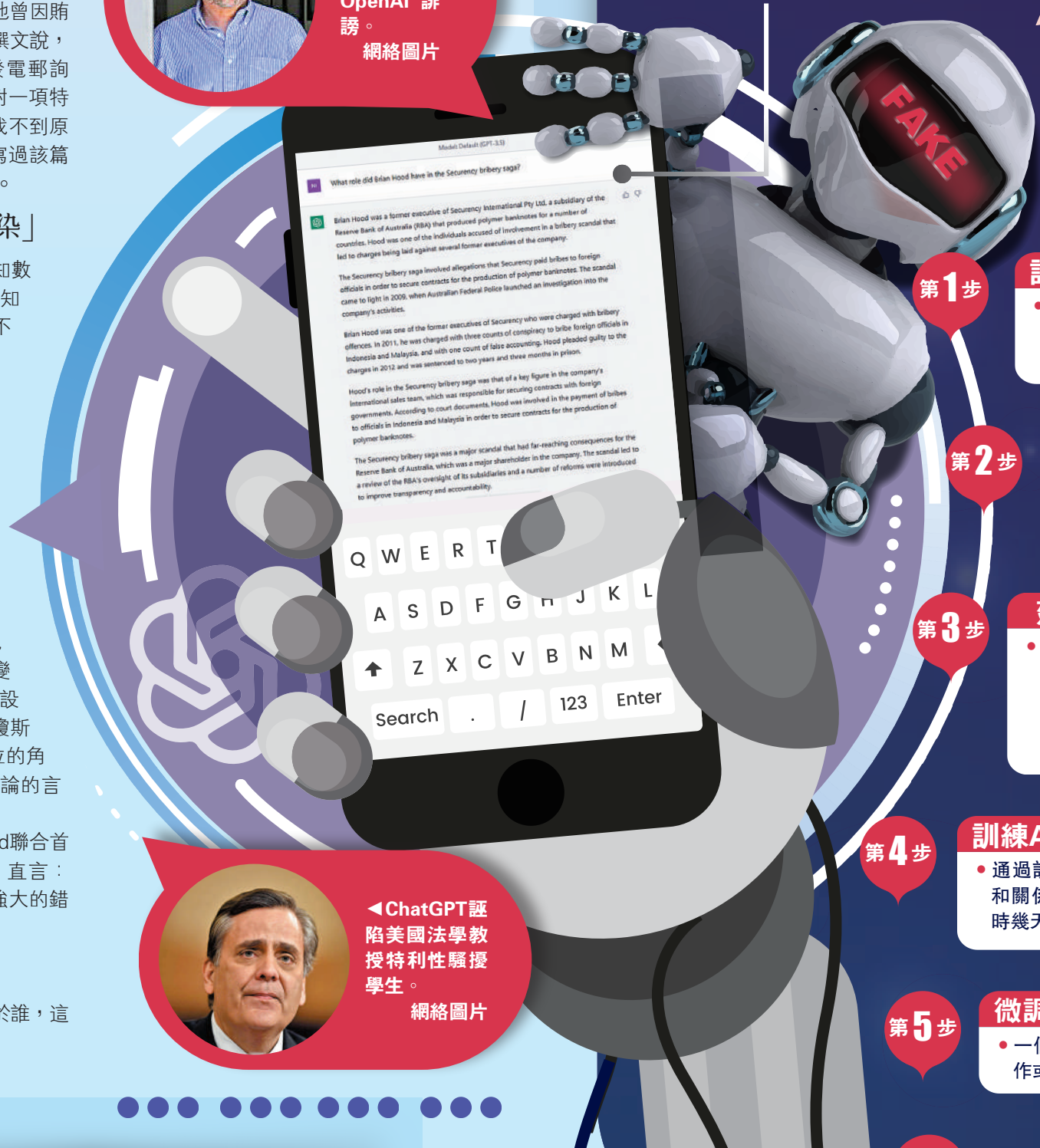
美國事實查核研究機構NewsGuard聯合首席執行官克羅維茲（Gordon Crovitz）直言：「這個工具將成為網絡上有史以來最強大的錯誤訊息工具。」

多國擬加強監管ChatGPT

當機器人編造假訊息時，責任在於誰，這還不清楚，因為目前還沒有AI亂講話導致公司被告的判例。美國《通訊端正法》第230條，保障科技公司可不必對平台上的第三方言論負責。聊天機器人從網絡上擷取文本、彙整後傳達給使用者，這是否屬於第三方言論，也很難判定，因此不確定科技公司能否用這條法規自保。

多國近期已收緊對ChatGPT監管。意大利個人數據保護局（DPA）宣布從3月31日起禁止使用ChatGPT，德國、法國、愛爾蘭等國家已開始準備效仿意大利的做法。加拿大聯邦隱私監管機構已經對OpenAI展開調查，因為該公司涉嫌「未經同意收集、使用和披露個人信息」。

◀澳洲官員胡德擬控告OpenAI 誹謗。 網絡圖片



ChatGPT杜撰的胡德行賄入獄假新聞。 網絡圖片

◀ChatGPT誣陷美國法學教授特利性騷擾學生。 網絡圖片

多方限制 ChatGPT

◀多國進一步對AI發展作出規管。 網絡圖片

- **意大利**政府以違規搜集私隱為由，從3月31日起暫時禁止使用ChatGPT，成為西方國家首例。德國也在考慮是否效仿意大利。OpenAI已向意大利當局提交解決方案文件。
- **歐盟**近期正在討論人工智能法案，如何限制高風險AI產品的新規例，如ChatGPT。美國總統拜登4日表示，AI是否危險有待觀察，強調科技公司有責任推出產品前確保安全，籲國會盡快通過私隱法。
- **韓國**三星電子疑似因員工使用ChatGPT，洩露公司機密，三星籲員工不得分享機密資訊。美國部分律所、摩根大通銀行、Verizon通訊等已限制連接ChatGPT。
- **美歐**多所大學已禁止學生使用ChatGPT，紐約市教育局禁止當地的公立學校電腦以及網絡使用ChatGPT。

大公報整理

ChatGPT如何工作：

- ChatGPT的工作原理是大型語言模型（large language model，簡稱LLM），是AI領域相對較新的訓練模型，約在5年前首次出現，如今已可以撰寫各種文章。

第1步 設定目標
- AI系統須預先設定目標函數，大多數LLM模型的基本目標函數為：給定一個文本序列，猜測接下來的內容。

第2步 收集大量數據
- 大量收集訓練數據，ChatGPT等通常從互聯網上搜集數十億個頁面作為數據庫，如博客文章、推文、維基百科和新聞。

第3步 建立神經網絡，組裝「大腦」
- 數據被拆分成標記單元輸入模型，可是單詞、短語或單個字符。接下來組裝人工智能的「大腦」：即AI的神經網絡系統。這是一個由相互連接的節點（或「神經元」）組成的複雜網絡，用於處理和存儲信息。

第4步 訓練AI「大腦」
- 通過訓練，該AI模型學會分析數據，識別不同模式和關係，學會如何構建有意義的信息。相關訓練耗時幾天甚至幾周，耗費巨大的計算能力。

第5步 微調模型
- 一個大型語言模型被訓練出來，需要為特定的工作或領域進行校準，通常由人類進行微調。

第6步 上線啟動

來源：《紐約時報》

醫生：患者勿依賴ChatGPT看病

【大公報訊】據《每日郵報》、商業內幕報道：聊天機器人ChatGPT功能強大，相較於通過搜索引擎在海量結果中尋找答案，能夠給出回答的ChatGPT似乎有不少用戶將其視爲「醫生」求醫問診。有醫生警告，不要使用ChatGPT獲取健康資訊，因爲研究發現ChatGPT提供的信息並非全部正確，甚至會編造虛假信息。

美國馬里蘭大學醫學院的研究人員要求ChatGPT回答25個與乳腺癌篩查建議相關的問題，且每個問題都讓ChatGPT回答了三遍，然後由三位乳房X光攝影術專業醫生分析其回答。結果ChatGPT正確回答了有關乳腺癌症狀、哪些人處於危險中以及X光檢查的費用、年齡和頻率建議的問題，其88%的答案都是恰當的。研究作者之一的Paul Yi博士稱，他們發現ChatGPT有時會編造虛假的期刊文章，或僞造健康機構來支持其說法，「用戶應該依靠醫生而不是ChatGPT來尋求建議。」

不過，也有醫生指出，ChatGPT的升級版GPT-4已成功通過了美國執業醫師資格考試（USMLE），在臨床診症已相當準確，未來對於輔助診症，還是有相當大的用處。哈佛大學計算機專家及內科醫生Isaac Kohane近期用GPT-4測試美國醫生資格考的題目測試，發現九成的答案都正確。對於發生率10萬分之1的罕見病先天性腎上腺增生症，GPT-4更是在幾秒內就能診斷出來。

也有研究指出，ChatGPT援引的醫學資料來源過於單一，加上AI不能進行臨床判斷也毋須擔負醫護道德責任，患病後不要依賴ChatGPT診症，必須要看醫生。

▲醫生警告患者向ChatGPT諮詢健康事宜應謹慎。 路透社

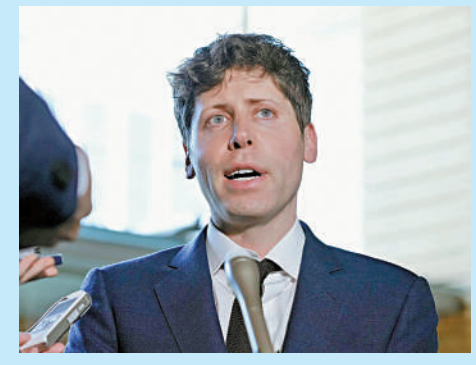
科企濫用內容訓練AI 出版商索要版權費

【大公報訊】據《華爾街日報》報道：隨着ChatGPT等AI聊天機器人走紅，目前全球多家傳媒及出版公司已開始審查旗下產品內容，在多大程度上被用於「訓練」ChatGPT等AI，並考慮採取法律行動，以便向AI研發商獲得相關內容的版權費用。

傳媒倡議組織「新聞媒體聯盟」執行副總裁兼首席法律顧問科菲表示，出版商應該知道如何獲得補償，「我們的內容是真正的人類辛勤勞動的成果，卻不斷被用來為其他人創造收入，我們必須得到補償。」

雖然美國和歐盟等地法律設有「合理使用」條款，允許個人和公司在特定情況下未經許可使用版權資料。不過出版商普遍認為，科企海量使用版權內容訓練AI，是一種「濫用特權」的行為。有出版業亦擔心，聊天機器人等改變互聯網搜索生態，將直接帶走本屬於其網站的流量和收入。

OpenAI行政總裁阿爾特曼表示，他們在合理使用方面做了很多工作，並願意為某些領域的優質數據支付高昂費用。據悉，美國社交平台Reddit已與微軟進行磋商。2月，在線圖庫Getty Images也對AI圖片生成軟件Stability AI提出起訴。此外，擁有《華爾街日報》、《紐約郵報》、《巴倫周刊》等美國新聞集團亦準備採取行動索償版權費。



◀OpenAI行政總裁阿爾特曼。 網絡圖片