AI有「幻覺」時常會「說謊」

專家：防止AI犯錯 人類干預不可缺

AI 利與弊（上）

人工智能（AI）在今年爆紅，這股由聊天機器人ChatGPT引發的風潮，在為人們帶來工作學習方便的同時，也讓外界認識到其潛在的不足和破壞力。不少用戶在使用ChatGPT等工具時，都會發現人工智能其實會「說謊」和捏造事實，絕非萬無一失。這種現象被稱為「AI幻覺」（AI Hallucination），其生成的答案真假難辨，甚至可以騙過專業人士的慧眼。專家指出，人類仍需在應用中干預AI的行為，包括在面對AI生成的信息時應格外注重事實查核，防止「AI幻覺」混淆視聽。《大公報》今起推出系列專題，多角度探討人工智能的利與弊。

大公報記者 楊金宇

近日，英媒曝出美國得州的特斯拉工廠曾於2021年發生流血事故，工廠內的機器人發生故障，「襲擊」了一名工程師致其受傷，成為「AI可傷人」的最新例證。同為AI的大型語言模型（LLM），即ChatGPT背後所使用的技術，亦有發生故障的可能，即系統產生「AI幻覺」。該情況發生時，AI會根據使用者的提示產生虛假的捏造信息，可能會對原文增刪、竄改，甚至杜撰內容，並將這些存在謬誤的信息「一本正經」地作為答案傳回給使用者。

就原理而言，大型語言模型接受了大量資料的訓練，它們利用這些知識來「理解」使用者的提示，並產生新內容。AI聊天機器人本質上是在預測句子中最有可能出現的詞語，因此它們有時會產生聽起來正確但實際上十分荒謬的答案。

AI為答問題「不擇手段」

香港浸會大學高級講師、AI和數字媒體課程副主任保羅・門戈尼博士接受採訪時指出，儘管AI用於構建答案的信息可能均為真實，但多個真實條件聯繫起來構建出的答案未必會符合邏輯。

對於AI幻覺的可能成因，門戈尼通俗地解釋稱，由於AI的設定限制了其對於用戶提出的問題「不能不知道」，因此會造成其想盡辦法回答用戶的問題，即使提供的答案存在謬誤，產生如AI幻覺的行為。

目前對於AI幻覺產生的錯誤信息，仍無準確有效的辨別方法，而這些錯誤信息甚至能夠蒙蔽住專業人士的雙眼。11月，一名巴西法官被發現使用ChatGPT撰寫的裁決出現多處法律案例上的錯誤。美國紐約一名執業30年的律師在5月份被發現在法庭文件中引用的多宗案例及文章內容均為虛假，他承認曾依賴ChatGPT搜尋資料，並已反覆向AI確認這些資料的真實性。

OpenAI和谷歌都警告用戶，AI聊天機器人可能會犯錯誤，並建議用戶在使用時仔細檢查其生成的答案。

勿依賴AI生成內容

門戈尼表示，AI如同人一般會犯錯，且其犯錯是很正常且相對常見的。但其亦同人一般會從錯誤中學習，從而進步。他指出，通過以更多數據對AI進行更多的訓練，將有助於減少AI出現幻覺或出錯的幾率。這也是訓練AI模型的核心，讓其在不斷犯錯中通過其神經網絡不斷學習改進，從而減少未來犯錯的幾率。

門戈尼稱，目前階段，AI對於人類最大的幫助仍是協助進行「頭腦風暴」，為人類提供思路與基礎。他指出，AI背後大量的數據能夠使其生成更多更好的想法，但是由於目前AI的邏輯能力仍然相對有限，其生成的一些想法可能脫離現實或無法實現，故而仍需人類承擔最終決策者的角色。

至於為何人類仍需在應用中干預AI的行為，門戈尼認為，在技術層面上，由於訓練AI的算法及原理限制，AI的行為極可能存在偏見；而在道德層面上，AI則無法為其行為付出代價，故而AI不可能完全取代人類。不過他也指出，不能因AI存在犯錯的可能性就忽視其巨大的創意潛力，因噎廢食，拒絕科技進步。

由於當下使用AI生成內容愈發普遍，門戈尼建議人們在使用AI的過程中，切勿對AI生成的內容產生「依賴感」。他還提醒普通民眾，應時刻秉持懷疑精神。在接觸AI或疑似AI生成的信息時，不應盡信自己所見所聽，主動進行事實查證，以人類的身份成為AI生成信息的最後一道門檻，「這也是為何人類在AI行為中不可或缺的原因之一。」

「AI幻覺」（AI Hallucination）

指AI的一種自信反應，其在回答相應問題時虛構事實，輸出看似有理有據，實際虛假或荒謬的答案，與心理學上幻覺的定義相似。對於非專業人士，可能難以判斷AI給出的答案是否符合事實，有可能成為假信息的「受害者」。

AI為何有「幻覺」？

「AI幻覺」時間表

- 1950–1956 AI誕生
- 2000 「AI幻覺」首次作為概念出現在研究文章中
- 2018 谷歌DeepMind研究人員公開提出「AI幻覺」的概念
- 2022 ChatGPT發布，「AI幻覺」一詞的使用量達到高峰
- 2023 英國劍橋詞典將「AI幻覺」選為年度代表詞

資料來源：Tidio網站

►香港浸會大學高級講師、AI和數字媒體課程副主任保羅・門戈尼博士。

不同AI模型出現「幻覺」幾率

OpenAI
GPT-4： 3%
GPT-3.5： 3.5%

Meta
Llama 2 70B： 5.1%
Llama 2 7B： 5.6%
Llama 2 13B： 5.9%

Anthropic
Claude 2： 8.5%

谷歌
Google Palm： 12.1%
Google Palm-Chat： 27.2%

數據來源：Medium網站

科企研對策消除「幻覺」

【大公報訊】事實上，AI幻覺一詞早就有相關學術文章討論，今年因為ChatGPT大熱，才開始真正引發關注和熱議。與此同時，AI出現幻覺可能產生嚴重的後果，科企們也在努力想對策消除這種「幻覺」。

今年2月，谷歌（Google）首次發布聊天機器人Bard之際，就因為宣傳視頻在回答問題時「張冠李戴」，導致谷歌的市值蒸發1000億美元。由谷歌前員工創立的新創公司Vectara研究估計，即使在對發生AI幻覺有所防備的情況下，ChatGPT也至少會有3%產生幻覺的幾率，而谷歌用於社交平台AI的Palm-Chat模型產生幻覺的幾率則高達27.2%。

各大科企同時也在研究減少幻覺的方法。谷歌表示，減少幻覺發生的一種方法是通過用戶回饋來改善AI：如果Bard產生的答案不準確，用戶應點擊「反對」按鈕並描述答案錯誤的原因，以便其可以學習和改進；OpenAI則實施了一項名為「流程監督」的策略。通過這種方法，AI模型會獎勵自己使用正確的推理來得出輸出，從而訓練模型產生人類認可的思想鏈。

OpenAI的mathgen研究員科布表示，檢測並減輕模型的邏輯錯誤或幻覺，是建構一致的通用AI的關鍵一步。

哪種AI容易出現「幻覺」？

AI出現「幻覺」，目前主要是針對大型語言模型（LLM）。這種系統可以編寫出合理的複雜文字，並生成圖片、影像和電腦代碼等，也是聊天機器人ChatGPT、圖像生產軟件Midjourney的基礎。

大型語言模型的訓練原理，是系統經過海量文本訓練進行深度學習，從而預測不同詞彙可能出現的序列。大型語言模型出現「幻覺」，很可能是因為訓練數據和模型架構而產生，具體原因暫時未知，且難以追溯源頭。

目前的研究顯示，以下情況或導致「AI幻覺」產生：

1. 過時、低質量、存在矛盾或傾向性的訓練數據在訓練過程中會誤導AI；
2. 當AI模型與訓練數據過度匹配時，可能導致其無法適應新數據；
3. AI模型亦可能不能充分理解上下文，從而產生脫離情境的虛假結果。

目前科學界尚未完全了解「AI幻覺」，減少其出現的研究和改進仍在進行中。例如，通過採取保證訓練數據的時效性與高質量；根據人類反饋改進AI算法；引入自動糾錯機制等措施，可以緩解「AI幻覺」的出現。

如何防止及減少「AI幻覺」產生？

提問更詳細

ChatGPT等聊天機器人背後的模型，需要適當的上下文才能產生準確的結果，否則其輸出將相當不可預測。用戶提問時須詳盡解釋需求，給出適當提示詞，令AI對問題有更全面的了解。用戶亦可使用特定的數據和來源來引導AI。

限制回應範圍

在使用AI時，模棱兩可的問題可能會被誤解，並增加產生「幻覺」的幾率。用戶應向AI設定其答案的範圍，提出有限回應範圍的選擇式問題，而非開放式的問題。

指定資料來源

用戶可以指定AI在特定來源尋找信息，確保AI使用經過驗證的來源，而非讓其從網絡上隨意獲取信息，能夠顯著減少產生「幻覺」的機會。

為AI分配角色

在提問時可為AI分配角色，如加入「你是某方面的專家」等提示詞，有助於為AI提供更多背景信息，並影響其生成答案的風格。由於AI模型將站在專家的立場上回覆，該行為還能夠提高事實的準確性。

交叉查證

對AI產生的答案，用戶千萬不能百分百照單全收，要通過常理進行判斷，或者搜索其他資料查證，難以把握的專業性問題，須諮詢相關專業人士。

「AI幻覺」造假實例

張冠李戴

- 2月，谷歌的聊天機器人Bard在回答問題時稱，第一張太陽系之外的星球照片是由韋伯太空望遠鏡拍攝到的，但實際上是由智利的甚大望遠鏡拍到的。

「冤假錯案」

- 3月，美國紐約州一名律師在替客戶撰寫案件摘要時，利用ChatGPT整理相關判決，後被發現其援引的多宗案例及文章內容均為杜撰。
- 4月，美國喬治城大學法學教授特利撰文，稱自己在3月底收到郵件稱，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。此事為子虛烏有。
- 11月，巴西一名法官使用撰寫的裁決中存在大量AI提供的錯誤信息，出錯裁決所引用的法院案件及法律先例中的某些細節，以及做出先前判決的法院信息均為錯誤的。

假書可奪命

- 美國紐約眞菌學會發布緊急公告，稱亞馬遜上存在大量由ChatGPT撰寫的蘑菇類科普書籍包含錯誤信息，輕信誤食可能致命。

傳播偽科學

- 英國BBC 9月報道，YouTube上存在大量由AI生成，包含偽科學、陰謀論等錯誤資訊的「教育性質」影片被推送給兒童，相關頻道觀看量達到數百萬。

「AI幻覺」危害多

散播虛假或誤導性信息

「AI幻覺」生成的假信息容易擴散，在醫療保健、金融或法律服務的領域，可能造成嚴重的後果，影響個人健康或造成財務損失等。「AI幻覺」甚至影響選舉，微軟生成式AI助理Bing Copilot近期被發現回答歐洲國家選舉相關問題時出錯，還編撰候選人的醜聞。

為企業帶來法律風險

所有行業中，目前有三分之一的企業已經以某種形式使用AI。AI工具出現幻覺可能造成現實後果，令企業承擔法律責任。例如，在金融服務行業，不正確的數字結果影響巨大。

加劇偏見和不平等

受AI模型訓練數據和算法影響，AI本身就可能已經存在相當大的偏見，而「AI幻覺」可能進一步加劇這種情況。以人臉識別為例，由於AI學習樣本主要來自白人，導致少數族裔受害最深，甚至導致冤假錯案。

大公報整理

責任編輯：楊金宇 美術編輯：賴國良