AI的價值觀：機器的偏見從哪來？

刊出日期：2023/11/14 | 文字：陳宣予 | 責任編輯：葉妍希

全文共**3094**字，閱讀大約需要**6**分鐘

2022年11月，OpenAI向全球推出了被譽為世界上最佳的人工智慧（AI）聊天機器人：ChatGPT，將AI應用推向全新的突破，AI繪圖、AI影像、AI創作紛紛接踵而至。然而，隨著生成式AI的爆炸性發展，問題也跟著浮上檯面。本文特別訪問東吳大學社會系教授兼院系級AI研究中心執行長劉育成老師，邀請他和我們談談，在人工智慧的快速發展下，人們和機器如何開始互相影響？

AI大不同，它怎麼影響我的價值觀？

ChatGPT出世後在全球掀起了一波熱潮，從理工科學到人文社會領域，都在談論AI語言模型對人類社會帶來的改變。除了應用層面外，隨著ChatGPT的普及，各國也開始加緊腳步投入語言模型的製作進程。除了美國開發的ChatGPT外，韓國的HyperClova，中國的悟道，法國的Bloom也都紛紛上架。《天下雜誌》將此現象形容為「一場大型語言模型軍備戰」，然而，語言模型的掌握為何如此重要？AI問答機器人可能引發什麼樣的「戰爭」？

「文字本身其實就是文化內涵的展現，光是文字本身，其實就很有可能讓我們對某些事情的看法不太一樣。」

劉育成教授以中研院詞庫小組（CKIP）開發的繁體中文語言模型「CKIP-Llama-2-7b」為例。此模型的測試版被網友發現，它對於國家、政治相關的提問會採用中國立場做答覆，包括國慶日是10月1日、領導人是習近平、開發者來自中國等。

「他們（開發者）是用什麼資料去訓練呢？他們的資料主要來自於中國的幾個資料庫，然後直接把簡體轉成繁體，再把它餵給他們的大型語言模型。這件事情一出來之後就被罵翻了，因為如果你今天只是把簡轉繁，那其實中國的用詞、他們喜歡的用語，事實上都沒有改變。」

劉育成教授提到，語言模型是根據現有資料建構而成，因此，人們給予什麼樣的資料去訓練，機器就會給予什麼樣的答覆，且這些回覆對我們的影響可能是無意識的。劉育成教授提醒，因為我們無法確機器背後是使用什麼語言模型在訓練，所以這些回答隱含的文化價值觀會無形的影響我們的認知。

▲ 劉育成教授提醒，AI語言模型可能會無意識影響人的價值觀。（照片來源 / 周芝辰攝）

AI語言模型對使用者價值觀的影響來自於它背後的資料庫，而資料來自於人類，不同種族、國籍的人都會握有其特定立場。即便立場並無對錯，但這些淺移默化的影響卻有可能被政治所利用，「大型語言模型軍備戰」也因此展開。但除了文化差異造成的政治攻防外，「人類本身」所持有的一切又是如何被AI過濾吸收？帶來了什麼樣的結果？

看不見的角落，掌握AI的回答的那雙手

早在多年前，各大公司就曾經嘗試過將AI聊天機器人推廣至社群上。「Tay」是一款由微軟團隊開發的聊天機器人，在2016年時於各大社交平台上線。Tay的目標群眾為社群媒體上的年輕族群，主打陪伴用戶輕鬆、愉快的聊天，並在與網友的互動中不斷學習。然而，不到一天的時間，Tay就因為說出了偏激的種族歧視言論而被緊急下架。但與前文中提到的中研院CKIP-Llama-2-7b不同，Tay的出錯不在於資料庫的選用偏誤，而是學習了網友刻意輸入的偏激言論。

▲ Tay機器人在推特上與網友交流，卻說出爭議言論。（圖片來源 / TayTweets）

「現在那幾個大型的語言模型公司掌握的就是參數的調整。他今天把這個東西開發出來之後，他只要透過參數的調整，就可以（決定）讓這個語言模型怎麼樣去回答這些問題。」

若將Tay與ChatGPT做比較，可以發現ChatGPT傾向中立的回答，而Tay則會更全盤的去消化、輸出得到的訊息。劉育成教授說明，這背後反映的即是參數的調整，開發者可以選擇過濾掉某些詞彙，也可以選擇讓AI自由發展。因此，看似自動又中立的AI系統，其實背後都蘊含了開發者的預設狀態，於是只要掌握參數調整的權力，就能掌控AI回覆的傾向。然而，這樣的權力掌握在極少數人手中，在公開透明的制度尚未實現前，人們在使用上的警覺就更為重要。

Tay的緊急下架，反映了大眾對於AI機器人的期待是「良善」的，盼望AI能對社會有正面影響。同時也顯示了「惡」即使是人性的一部份，在人工智慧的運用上仍希望能儘量避免。但即便開發者能設法讓演算法本身中立，可人的既有立場是否也會成為一種偏見？

人工智慧，人的偏見

隨著AI對人類的影響逐漸擴大，批判AI偏見的思考也同時產生。Netflix紀錄片《編碼偏見》中，提出了人臉辨識系統對膚色判讀的落差：片中科學家Joy Buolamwini起初是想開發一款具有人臉辨識功能的鏡子，卻意外發現，人臉辨識系統判讀淺色人種的準確度高於深色人種；辨識男人的準確度高於女人，開啟了她對於AI偏見的研究，揭露了系統存在的偏差。

▲ Netflix紀錄片《編碼偏見》揭露人臉辨識系統的偏差。（圖片來源 / Youtube）

《路透社》也曾報導，全球最大網路電商亞馬遜（Amazon）過去曾使用AI來篩選入職者履歷，其中卻存在性別歧視。由於公司提供給AI的資料本身存在性別比例不均的狀況，因此AI在計算這些資料後，得出帶有歧視性的結果。此事件凸顯AI演算法會吸收人類提供的資料的偏差，並將其加入自動化的計算，得到有偏見的結果。

劉育成教授提到，不單是大數據，就如同統計的資料，都會需要進一步處理才能消除這些偏誤。

「數據本身就有偏見，那數據為什麼會有偏見？就是因為數據訓練的數據 都來自人。那人就是有偏見的，所以數據本身它就會帶有一定的偏見。」

人與機器，人與未來

在人類同樣帶有偏見的狀態下，對於AI偏見，人們想防範的究竟是什麼？劉育成教授以亞馬遜用AI應徵員工為例，有人會因為AI的性別偏見，而失去面試機會。AI法規即是希望能解決這樣的問題：確保AI偏見不會對真實世界造成影響。

「AI的法規其實就是要去處理它可能對現實帶來的影響。因為偏見是態度性的，但它會怎麼影響真實世界？所以現在很多深層次AI的規範都會去連結到說，我們要怎麼樣能夠確保它對真實世界不會帶來實質的影響。」

人類與AI的連結環環相扣：人們將問題交給AI解決，但AI帶來的問題卻只有人類能解。人在探討人工智慧帶來的隱憂和影響時，最終還是會回到「人類」本身，正因為偏見難以避免，在談論人工智慧時，人與機器的差異似乎更加明確。

「人的思考跟演算法不一樣，我們有反思的能力。今天也許你會說出一個帶有偏見的話，可是你會反思說這個可不可以做、這個好不好，我們就會在這過程裡面就會去修正我們的想法跟行為。可是對演算法來講，它不會。」

劉育成教授說明，AI的運算倚靠資料庫，如果這些資料存在著偏見，AI就必須要「學習一個什麼」才能修正它，那如果機器沒有學習到這些內容，這個偏見就可能會不斷地重複。

人會反思自我行為，而這也是機器和人類最大的不同之處。提及「人的偏見」，多數人會將其視為負面影響，但與AI相比，人之所以能夠不斷在異中求同，尋找最佳解，就是因為人具有道德感和反思能力。因此，當人們在運用AI工具時，若能秉持懷疑、批判演算法偏見的心態，或許能更好的在使用上取得平衡。

人人都能使用的人工智慧，人人都準備好了嗎？

參考資料：

Amazon scraps secret AI recruiting tool that showed bias against women

台版GPT出爐！ChatGPT都吃簡體資料，他們打造全球第一個繁中模型

聊天機器人Tay 上線一天就「學壞了」，成了一個種族歧視者

人工智慧會不會