莊庭瑞
中央研究院資訊科學研究所副研究員，
合聘於資訊科技創新研究中心以及
人文社會科學研究中心。

資料為體，系統為用

受資料左右的人工智慧系統，如何左右身為使用者的你我？

撰文／莊庭瑞

「機器學習」已是常見的用語，多數人都不陌生。簡單地說，機器學習使用自動化的方法分析給定的資料，生成高效率的運算模型，用來判別之後同類型的資料。用來訓練模型的資料越多，得到的模型也就越精細。當然，訓練資料的特質，影響了生成模型的特性。因為資料缺失導致模型偏差的情形也時有所見，卻很難避免。以下舉個例子。

假設手邊有大量貓和狗的照片，其中出現的狗都是黑狗兄，貓則有花貓、白貓、但就是沒有黑貓。只用這批照片訓練出來的貓狗辨別模型，看到毛茸茸一團可愛的黑麻糬，十有八九會認為那是狗。

上述例子當個笑話，應該無傷大雅。這類失誤若是出現在人臉辨識系統，當事人可就要翻臉了。2018年由博蘭威尼（Joy Buolamwini）和葛布若（Timnit Gebru）兩位非裔女性計算科學家合著的論文指出，市面上三款分別由微軟、IBM以及中國某廠商製作的人臉辨識系統，遇到黑皮膚的人頭影像，常錯把女當男、或是男當女，有些系統的錯誤率可高達30%。但是這三家系統卻都能精準辨別白皮膚頭像的性別。其中緣故可以想見，應該是用來訓練的照片中白人遠多於黑人，甚至所收錄的黑人頭像原本就有性別標示錯誤的情形。

值得一提的是這兩位研究人員的巧思：用來測試三家系統的資料，取自冰島、瑞典、芬蘭、南非、塞內加爾、盧安達的國會議員頭像和性別。這些都是公開資訊。六國共1270位代表組成的測試資料集，在膚色與性別的組成，比當時市面上的測試資料集，均衡許多。

機器學習所用到的訓練資料和測試資料，必須符合之後會遇到的資料，否則再多資料訓練出來的模型，還是不符需求。這是訓練資料的涵蓋性議題。另一方面，訓練資料因其取材來源，也會帶進成見，這在自然語言處理尤其可能。在資訊爆炸的網路時代，語料庫來自四處爬梳的網頁，難免包括不雅詞語和偏見，這需要費心過濾，否則訓練出來的模型也會說髒話、帶有歧視。有成見的網頁資料訓練出有成見的模型，這模型若用來產生更多帶成見的文句、散佈在網路，就成為惡性循環。

針對自然語言處理，美國西雅圖華盛頓大學的班德（Emily M. Bender）等三位教授，提出「資料陳述」（Data Statement）的撰寫指引，用來描述語文資料集的基本資訊，包括蒐集緣由與材料來源，希望有助於減緩資料集潛藏（不可避免的）內在偏誤所帶來的傷害。日常語句承載了社會成見，例如大眾在媒體常看到「亞裔美國孩子會念書」、「非裔美國人會運動」。有些成見容易察覺，有些則相當細微，大眾罕有感覺。

例如，「兩位非裔女性計算科學家合著的論文指出」這句子真的有必要把「非裔」與「女性」放在「計算科學家」前頭做為修飾嗎？為什麼需要指出這兩位研究人員的族裔與性別？是認為計算科學家少有非裔女性，所以需要指出並強調嗎？這是否就是成見？

去年初，班德與葛布若領銜發表論文〈隨機學語鸚鵡的危險：語言模型會太大了嗎？🦜〉（On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜）。是的，你沒看錯，那篇論文的標題的確以鸚鵡的表情符號做為結尾。對於使用（品質堪慮的）超巨量的語料庫來訓練語言模型，此文提出許多批評與建議。其一，這些模型固化了原有語料中的成見，也可能因語料的不當過濾，模型難以符合弱勢族群的需求；其二，訓練這些超巨量模型需要大量耗能，產生可觀的碳足跡。

原本任職Google的葛布若，為了這篇論文初稿槓上公司而遭解職（一說自行辭職），同在Google任職的其他作者則埋名以對。一年後風波平息，但議論持續。

網路普及帶來資料大爆發，搭配低廉的運算及儲存成本，為人工智慧系統創造突破性的進展，當這些系統反過來影響眾人對世界的認知時，我們也該對訓練人工智慧系統的資料組成與限制，有更多的思考。SA

影像來源：莊庭瑞