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。

機器學習與資料倫理

人工智慧系統生成的影像，是否用了你的作品？

撰文／莊庭瑞

近年來資訊巨擘熱中於開發「文轉圖」人工智慧（AI）系統，以機器學習方法處理數十億張配有圖說的影像，產出方便使用的運算模型。使用者輸入簡短文字描述，系統隨即產出相稱的新奇圖片。這類AI系統基本上是類神經網路，由上千億參數連接組合而成，多如繁星的參數值經由大量計算、處理巨量資料後調適而成。

2022年發佈的這類文轉圖系統包括新創公司OpenAI（由微軟資助）推出的DALL-E 2以及Google所研發的Imagen。這兩項系統需要申請且僅能透過網際網路使用，使用者透過瀏覽器輸入指示，生成的圖片再經由網路傳回。運算模型是商業機密，置於雲端有如黑盒子。許多人好奇於這些黑盒子的能耐，於是跟這些公司申請使用許可，透過網路進行試驗。

德國慕尼黑大學以及新創公司Stability AI、Runway在同年8月共同推出的文轉圖系統Stable Diffusion，引發了一陣騷動，因為研究團隊把運算模型也公開了。儘管模型還是黑盒子，這次卻能下載到個人的電腦自行使用，更可以調適為新模型，或是開發新的資料介接和使用介面。應用情境頓時百家齊鳴，但也觸發新的議題。在網站Reddit有位發文者說，已使用該運算模型，套進插畫家孟格（Hollie Mengert）的作品，做出了可以自動產生迪士尼風格插畫的模型，供眾人試驗。孟格經朋友通知後，才知道這回事。她心中感受如何，不難想見。

此外，訓練資料本身的偏差會帶進機器學習產出的模型，例如生成的醫師照是白人遠多於黑人。經由網路介面使用的文轉圖系統，在文字輸入端或是圖片輸出端，提供服務的公司經常會再過濾選取，希望去除資料內在的偏見，或是以服務條款要求使用者不能惡搞。這些補救措施跟應用情境高度相關，實在掛一漏萬。來自研究倫理上的指引，尤其在資料使用方面就分外重要。

Stable Diffusion團隊不僅公開運算模型，也把程式以及用來訓練的所有影像檔案一併釋出給公眾使用檢視。研究專案產出的程式碼通常以開放源碼（open source）授權條款釋出，常見的授權條款，例如通用公眾授權條款（GPL）或美國麻省理工學院（MIT）授權條款，讓使用者可以自由修改程式碼，而且不限制（修改後）程式的使用目的。開放源碼的授權方式可適合用來釋出運算模型？這問題引發許多討論。首先，運算模型是資料還是程式？見仁見智，看法不一。至於像是「不能用來惡搞！」這種使用目的上的限制，在常見的開放源碼授權條款上也幾乎沒有列出。

基於這些考量，Stable Diffusion另以特別授權條款釋出運算模型，雖然允許使用者下載、複製、改用模型，但加上許多限制，其中包括不得違反當地與國際法律、不能企圖用於剝削或傷害兒童、不可用於生成或傳播不實資訊或內容等。這類加上使用限制的運算模型授權條款，一般稱為負責任的人工智慧條款（RAIL）。

同樣受到關注、也在2022年發表的大型語言模型BLOOM應是首先採用RAIL授權條款的大型研究計畫。這計畫有上千位參與者，來自70多個國家，強調以開放協作方式推展AI的進程，並把語料庫、程式碼、運算模型都釋出給公眾使用。計畫採用的語料庫包括了46種自然語言以及13種程式語言。

本文附上用Stable Diffusion生成的一張圖片，筆者給系統的指示：Taipei 101 Van Gogh。讀者是否覺得這圖有模有樣？不過這不是梵谷畫的喔！SA

影像來源：莊庭瑞、Stable Diffusion