公平公正的AI，可能嗎？

highscope.ch.ntu.edu.tw/wordpress/

公平公正的AI，可能嗎？
編譯／台大資工系 黃柏瑋

在這逐漸自動化的世界裡，許多金融或政府機構已開始利用AI來審查貸款、社會福利、簽證，甚至於職缺應徵資格。但AI就像是一面鏡子，如實地反映訓練集中的資料，也難免透露出背後人類固有的偏好與社會不公，導致演算結果對弱勢族群的歧視，甚至因行政效率的提升，加速剝削的發生。「演算法的公平性與透明性」於是逐漸受到重視，歐盟更試圖將「可解釋AI」編入法律，要求所有「個人化自動決策」（Automated Individual Decision-Making）賦予用戶請求解釋的權利，並確保訓練數據的中立，不得讓任何用戶的權益因此受損。只是，如「黑盒子」般神秘又複雜的神經網路，其判斷依據可能就連開發者都理不清。

反事實解釋

事實上，演算法的透明性並沒有一個明確的標準，銀行真有必要公開貸款系統的程式碼以達到全然的透明嗎？企業又應該公布多少比例的資訊，才能讓應徵者足夠瞭解篩選演算法的運作？

「（直接）揭露公司的程式碼是下下策，如果我們只考慮真正影響決定的因素時，其實不必明白機器究竟用了哪些公式。畢竟，人們更想要知道評判的依據是什麼。」牛津大學數據倫理與網路規範教授Sandra Watcher說道。反事實解釋（counterfactual explanation）：令系統直接告知用戶翻轉原先判決、達到理想結果的最快方法，反而更能滿足大眾對透明化系統的期望。

以借貸資格的評判為例，一個透明化系統應能夠向用戶解釋：為什麼被拒絕貸款？如何才能翻轉原判決？或許是因為帳戶中的存款不夠，若儲蓄達到某個特定金額後，便可獲得貸款資格。與此同時，我們也可以利用這些解釋和建議，監控演算法的公平性：例如當系統向用戶提出悖離現實的解釋或要求，像變更膚色或返老還童等時，開發員便可依此修正機器的訓練

資料或演算法，保障用戶權利。

解鈴還需繫鈴人

只是，「反事實解釋」便能保證AI的公平性嗎？這之中其實也有模糊的空間，例如給予生活較富裕的、年紀較輕男性客戶相對輕鬆的標準和建議，忽略所謂「實質上的公平」（equity）。某些保險AI要求女性達到和男性一樣多的體脂，方可取得同等的健保費率；然而，男女的生理條件原本就生來不同，這樣的評判標準顯然並不合理。

此外，我們也很難完整定義所有「不公」的變數，以便演算法在做決斷時排除。舉例來說，以「性別」、「年紀」、「種族」為依據的演算法很容易被發現、偵錯並修正，但如「體力」、「郵遞區號」等具潛在歧視意味的判準，卻很難在第一時間被剔除。

事實上，一個再怎麼不偏不倚的AI，也無法導正一個本質不公的社會。根據報導指出，美國移民及海關執法局（U.S. Immigration and Customs Enforcement，ICE）曾經對一建議「是否拘留或釋放即將被驅逐的待審移民」的AI動過手腳，他們將「建議釋放」的選項由演算法中移除，雖移民官仍可選擇無視機器的建議另為判決，但亦導致被拘留的移民人數因此暴增。

演算法怎樣才算公平？並非純粹技術上的問題，更重要的是圍繞在演算法這項工具周遭的社會氛圍與公眾意識。

編譯來源

L. Matsakis, “What does a fair algorithm actually look like? ”, *WIRED*, 11 Oct 2018.

參考資料

C. Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable", Ch 6.1, Christophm.github.io, 2018.

(本文由教育部補助「AI報報–AI科普推廣計畫」執行團隊編譯)