標題：演算法公平嗎

關鍵字：演算法、公平性、反事實解釋

AI 就像是一面鏡子，除了可以模仿人類的思維之外，也能夠如實地反映出社會中難免存在的不公，導致演算法可能歧視某些弱勢族群，甚至隨著行事效率的提升，加速剝削他們的權利。更嚴重的是，AI 如同「黑盒子」般神秘又複雜，撲朔迷離的判斷依據可能就連開發者都理不清，這也代表人們將無法在逐漸自動化的世界裡，理解為何貸款資格被取消、福利被沒收、或者應徵工作失敗。因此，「演算法的公平性與透明性」逐漸受到重視，一般資料保護規範(GDPR)更直接將「可解釋的 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，並確保訓練數據中立，不得讓任何用戶的權益由此受損。

反事實解釋

事實上，演算法的透明性並沒有一個明訂的標準，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？

Sandra Watcher 曾經說過：「揭開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

知道評判的依據是甚麼。」而讓機器做出反事實解釋(counterfactual explanation)，並提供用戶達到理想結果的最快方法，便可以讓系統的分類標準更加透明化。

以借貸資格的評判系統為例，它應該要解釋為什麼這個用戶被拒絕貸款，也要提供能夠翻轉決定的方法。可能是因為帳戶中的存款不夠，而在達到某個特定金額的儲蓄之後，便有機會獲得貸款的資格。當然，我們也可以利用這些解釋和建議，監控演算法的公平性，假設系統向用戶提出悖離現實的要求，像是變更膚色或返老還童，開發員就必須修正機器的學習資料或演算法，好保障所有用戶族群的權利。

然而，一條案例可能會有超過一項解釋，究竟哪一個才是最好的呢？倘若他們都相當不錯，但是相距甚遠，那我們應該提拱哪些給用戶呢？這些問題目前沒有一個定論，可能套用在人力篩選的規則未必能被借貸公司接受。而且目前在這方面也缺乏具有普遍性的軟體開發，所幸它的操作簡單了然，還有一些線上網站利用「數據視覺化」協助探索模型的決策邊界，減輕開發者不少的負擔，也讓不會寫程式的人能夠參與分析過程。

偏私的根因

我們能單單從演算法提供的解釋就能斷言它的公平性嗎？或許它會有些投機，給予生活較富裕的、年紀較輕的或男性的客戶相對輕鬆的標準和建議，甚至也忽略掉所謂「實質上的公平」。像是某些健保 AI 要求女性達到和男性一樣多的體脂，以取得等值的健保費率；然而，男女的身體結構原本就生來不同，這樣的評判顯然有些脫離現實。

另外，僅有解釋的幫助，是無法讓我們得知演算法在判斷時最優先考量的變數，甚至我們也很難將所有「不公」的變數全部定義明白。舉例來說，以「性別」、「年紀」、「種族」作為依據的演算法很容易被發現並篩除，但我們卻很難挑出像是「體力」或「郵遞區號」等稍帶有歧視意味的分類。

事實上，一個再怎麼不偏不倚的 AI，也無法成功導正一個本質不公平的社會。根據報導指出，美國移民及海關執法局(ICE)曾經對一台建議「是否要拘留即將被驅逐的待審移民」的電腦動過手腳，他們幾乎將「建議釋放」的選項全部移除，若電腦還是一不小心「建議釋放」待審移民，工作人員也會選擇無視，導致被拘留的移民人數在人為操控之下暴增不少。

就如同 Madeleine Clare Elish 說過的：「這個關於『演算法怎樣才算公平？』的問題，要的不光只是技術上的解答。更重要的是，社會上的各種運作，該如何到位地使用演算法這項工具。」

參考資料：

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

2. Christoph Molnar, “Interpretable Machine Learning”, 2018, ch6.1.4