偏見如何形成？由人類到AI

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

偏見如何形成？由人類到AI
編譯／黃柏瑋

偏見，針對特定個體或族群產生難以事實或其他跡證撼動的既定印象，看似一種涉及高等認知能力、專屬於人類的社會現象。即使許多電腦演算法的公平性受到質疑，但多被歸因於訓練集中無法完全剔除的人類「偏見」所致，演算法最多只是忠實反映人類社會中既有的種族或性別問題。然而英國卡地夫大學與美國麻省理工學院共同主持的一篇研究，卻認為看似「不經大腦」的「有樣學樣」或許正是促使偏見形成的禍首。

團體內偏私 VS. 團體外偏見

研究使用超級電腦進行上千次的模擬：每一回，虛擬個體都需要決定是否將資源贈予與自身同群，或群外的其他個體。依據「演化博弈理論」（Evolutionary Game Theory），多數個體理所當然地傾向選擇與自己相似、同群的個體為受贈者，以確保短期內能獲得相應的回報。

然而就心理學角度而言，這樣的「團體內偏私」（in-group favoritism）是中性的現象，不涉及負面態度，也不足以構成歧視；相形之下，針對非我族類的忽視，乃至於輕蔑或競爭心理，所謂「團體外偏見」（out-group prejudice）才是我們應該關注的目標，只是兩者往往是相輔相成的。

地方性利他主義（parochial altruism）便是「團體內偏私」和「團體外偏見」共同演進的實例之一。通常發生在兩國交戰時，士兵們愛護自己的國家與戰友，讓保家衛國的意志更為堅定，同時也加深對敵人的憎恨，燃起剷除異己的勇氣和效率。

流動性

值得注意的是，偏見具有流動性，能隨著人與人之間的交流而流傳與演進，顯示文化在其中扮演著關鍵角色。除了受贈者過去的表現與名聲，施予者也會受到周遭其他人的選擇所影響。「排斥異己」久而久之成為可觀察的外顯特徵，反過來成為團體的象徵與凝聚力量。

在人工智慧的啟示

為了確保較高的短期利益，個體會複製其他人所採用、投資報酬率較高的選擇策略，而這樣的過程並不需要特別優越的認知能力。具有基本認知與自我控制能力，且依賴周遭夥伴提供資訊的智慧體，例如自動駕駛車輛和物聯網（IoT）裝置，便有可能形成「偏見」。

幸運的是，研究也發現團體內次族群的數量越多，可以抑制偏見的形成，避免任何一方成為被剝削的弱勢。當然，前提是這些次族群必須樂於與其他次族群互動。

編譯來源

Cardiff University, "Could AI robots develop prejudice on their own?", *Cardiff University News*, 2019.

參考資料

1. R. Whitaker, G. Colombo and D. Rand, " Indirect Reciprocity and the Evolution of Prejudicial Groups", *Scientific Reports*, vol. 8, no. 1, 2018.
2. T. Yamagishi and N. Mifune, "Parochial altruism: does it explain modern human group psychology?", *Current Opinion in Psychology*, vol. 7, pp. 39-43, 2016.

(本文由教育部補助「AI報報–AI科普推廣計畫」執行團隊編譯)