標題：AI 會有偏見？

關鍵字：偏見、進化博弈理論、內團體

前一陣子許多電腦演算法的公平性受到質疑，透過人類的歷史資料訓練模組，可能會有意無意地灌輸「偏見」，加上機器的自動化，AI 恐怕將加速惡化社會上的種種不公。而且近期研究指出，AI 在和其他個體互動的過程中，可以經由辨識、複製等行為同化彼此，推演出不同「小圈子」的邊界，對於同類和異類之間的「偏見」油然而生。

偏見的組成

進化博弈理論(Evolutionary game theory)提供一個強大的測試框架來觀察歧視行為的根據和演進。有研究顯示，在以標籤為底的模型下實行「給予和索取遊戲(game of give and take)」時，當接受者的標籤和施主相近，自發性互動的機率會提升。事實上，標籤可說是個體行為的策略準則，也是身分認同的一項依據，而後對於價值相近的成員給予較多關心和肯定，形成所謂的「內團體偏私(in-group favoritism)」。

然而就心理學角度而言，「內團體偏私」並不涉及負面態度，其實不足以構成帶有惡意的歧視行為；反觀，我們應該對團體外成員的交流投以更多討論。由於內團體成員往往對於團體外的生活無法感同身受，導致誤解和偏見的產生，甚

至出現輕視、競爭、反對等舉動，這現象稱為「外團體偏見(out-group prejudice)」。

地方型利他主義(parochial altruism)為「內團體偏私」和「外團體偏見」的一種實踐，兩者共同推演出戰爭上攸關勝負的歧視性行為，士兵們愛護自己的軍隊，讓意志更堅定，同時也加深對敵人的憎恨，燃起剷除異己的勇氣和效率。

在社會中的偏見具有流動性，會隨著人與人的交流經歷演化、尋求團體的平衡和穩定。通常來說，人類社會就如同標籤模型，但高智慧的人們擁有直接辨識標籤的能力，能夠從表情或行為的反應推出對方的態度。因此，歧視性的態度自然成了團體的分界，尤其對於外團體的種種偏見，可以更明確表現出彼此間不同的氣氛。

偏見的演進

要對特定的對象產生成見看似至少需要有正常的人類智商，但其實不然。科學報導(Scientific Reports)上有篇利用電腦模擬有偏見的人類或 AI 行為的研究，其中在「給予和索取遊戲(game of give and take)」中，受測者會根據對方的身分和自己的贈與策略行動，當然也考慮了對於他者的偏見程度。

Roger Whitaker 教授說道：「在經過上千上萬次的模擬之後，我們會發現 AI 的偏見會藉由學習新策略而有所變化。」為了得取更高的獲益，他們會從內團體或外團體的成員上複製投資報酬率較高的辦法，所屬的團體也可能在潛移默化中有些更動，而這些辨識和決定並不需要特別優越的認知能力，但這也代表能夠自我控制的裝置將容易受制於附近的夥伴，自動車和物聯網(IoT)就是兩個近來的例子。

此外，這份報告也發現較多的內團體互動會提升合作和歧視的機會，但開放性的學習卻能夠在合作的同時，降低偏見的發生。而不同種類的亞族群越多，也可以擊落偏見和歧視在社會中的地位，Whitaker 教授為此下了定論：「大量的亞族群可以讓不具偏見的團體互相合作，避免成為被剝削的弱勢。當然，這也只會發生在樂於和外團體互動的環境之中。」

參考資料：

1. Cardiff University, “Could AI robots develop prejudice on their own”, Science Daily, 06 Sep 2018

2. Roger M. Whitaker, Gualtiero B. Colombo, David G. Rand, “Indirect Reciprocity and

the Evolution of Prejudicial Groups", Scientific Reports, 05 Sep 2018

3. Toshio Yamagishi, Nobuhiro Mifune, "Parochial altruism: does it explain modern human group psychology", Science Direct, 2015