DOI:10.6129/CJP.201909_61(3).0002
中華心理學刊 民 108，61 卷，3 期，197-208
Chinese Journal of Psychology 2019, Vol.61, No.3, 197-208

理解使用監督式學習而潛在有偏誤的人工代理者：認知心理學與認知神經科學的觀點

黃從仁 [1, 2, 3, 4, 5, 6]
國立台灣大學心理學系 [1]
國立台灣大學神經生物與認知科學研究中心 [2]
國立台灣大學人工智慧與機器人研究中心 [3]
國立台灣大學計量理論與應用研究中心 [4]
科技部人工智慧技術暨全幅健康照護聯合研究中心 [5]
科技部人工智慧生技醫療創新研究中心 [6]

人工代理者雖然是機器，但和人類一樣常會做出具有偏誤的決策。本文討論人工代理者中常用的機器學習系統何時會學習去做偏誤決策，以及如何使用認知心理學與認知神經科學中發展出來的方法來瞭解其具有偏誤的決策歷程。具體而言，我們會闡述本質上是歸納推理的監督式機器學習如何導致如忽略少數團體等不透明的決策偏誤。接著，我們會視一個人工代理者如一位人類研究參與者，回顧文獻中如何透過腦部切除與影像遮蔽等認知科學中的神經與行為方法來揭露一個人工代理者的決策準則與傾向。在文末，我們會討論有偏誤的人工代理者對於社會的影響，並鼓勵認知科學家們一同來揭示並改正機器的各種偏誤。

關鍵詞：人工智慧、深度學習、認知心理學、認知神經科學、機器學習