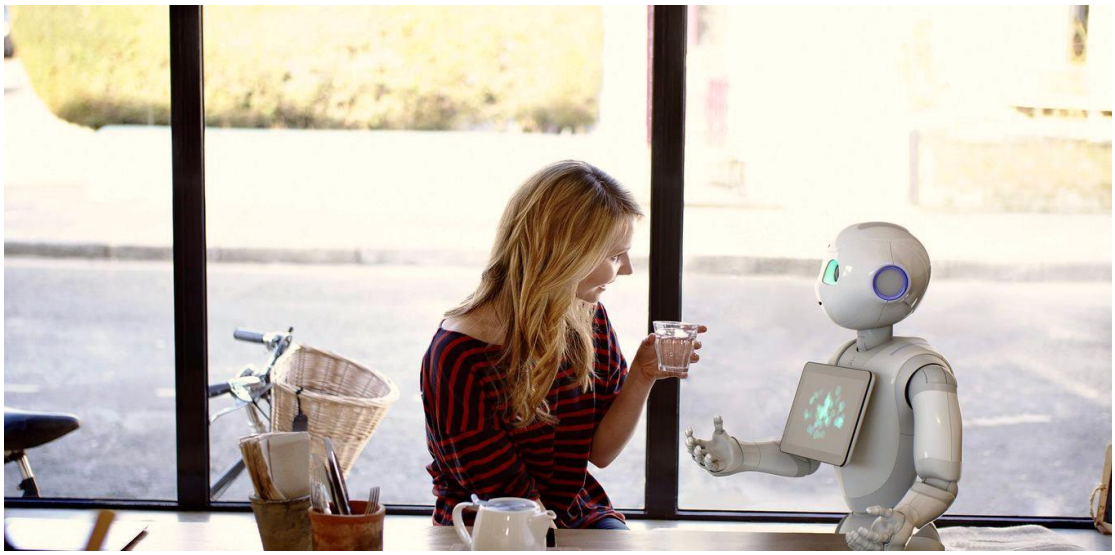
標題：探索 AI 的心智領域
關鍵字：心智理論、錯誤信念、ToMnet、情緒辨識



我們在跟 Siri 或是 Google Assistant 對話時，經常會因為他們的一無所知或答非所問而燃起無名火，那究竟為什麼人與智慧型助理的互動不能像人跟人之間的互動一樣順利呢？其實有部分是因為他們「理解他者」的能力不足，人工智慧若只有感知和思維是不夠的，沒有情感的引導，智慧系統還是無法發揮它應有的作用；但這也代表倘若人工智慧能夠理解周遭智能體的思維並做出適當的反應，便能將 AI 技術推往嶄新的發展階段，縮小人工智慧與人類智慧之間的鴻溝，讓人工智慧在這需要互動與溝通的社會上得到更有效的利用。

心智理論與錯誤信念

心智理論(Theory of mind，縮寫為 ToM)為心理學術語，泛指能夠理解自己以及周圍人類心理狀態的能力，這些心理狀態包括情緒、信仰、意圖、認知等等。心智理論這能力會依據動物種類、先天素質、生活經驗等因素而有所不同，例如相較於其他哺乳類動物，人類具有較高的心智理論能力，其中文學作家的能力又更為出色。根據研究，四到五歲的正常孩童即擁有基本的 ToM 能力，能夠成功理解他人的錯誤信念(false belief)與事實的差別，而 Sally-Anne 測驗為檢測此能力最常見的工具。測驗如下：Sally 和 Anne 同時坐在一個籃子與一個箱子的旁邊，Sally 將一顆紅球放入籃子後離開現場，接著 Anne 在保證搬運過程不被 Sally 發現的情況下把球從籃子移至箱子中，試問當 Sally 一回到這個空間時，會最先到哪裡找出紅球？如果測驗者認為 Sally 會到籃子裡找紅球，代表該測驗者能夠理解 Sally 因為沒有目擊搬運過程所導致與事實(球已被移到箱子裡)相歧的錯誤信念(認為紅球在籃子裡)，擁有 ToM 的基本能力；反之，若測驗者認為 Sally 會打開箱子找球，則代表測驗者無法以 Sally 的立場思考情況，ToM 能力有待加強。

機器心智理論網路—機器如何理解機器

在 AI 世界裡，類神經網路被稱為人工智慧三巨頭之一，是極具代表性的演算法，但其複雜程度超乎人類的理解能力，就好比黑盒子般令人難以參透，使得人們不好掌握問題的發展，以致於不敢信任人工智慧；而面對這種情況，有些專家認為「機器與人類看到的不一樣」，所以嘗試運用另一個人工智慧幫助人類理解原本的神經系統。

DeepMind 的研究人員 Neil Rabinowitz 及他的同事們受到了人類心智理論的啟發，設計一款機器心智理論網路(ToMnet)，並觀察它是如何理解其他人工智慧的行為。ToMnet 擁有三組神經網路：第一組網路會根據其他 AI 過去的舉動學習他們行為的傾向性，第二組網路會學習理解其他 AI 現有的信念(belief)，第三組網路會從前兩組網路獲得輸出並且依照情況預測其他 AI 接下來的舉措。

為了訓練 ToMnet 的理解能力，研究人員使被研究的 AI 在虛擬空間中移動，這些 AI 是會蒐集彩色盒子以賺取分數的簡單角色，而 ToMnet 將會從空間的上方觀察角色的行動；在一項測驗中有三種角色—看不見周遭環境的「盲人」、無法記得最近腳步的「健忘者」、以及看的見且記憶力良好的「正常人」，而每個角色取分的行為也所不同—盲人會靠牆前進、健忘者會接近離他最近的盒子、正常人則會運用策略安排蒐集盒子的順序以得高分。ToMnet 在經過一些訓練之後，能夠根據取分的行為辨別角色的種類，並且預測角色們未來的表現，這也代表著 ToMnet 可以理解角色們因為天生素質不同而做出的各種行為。

另外，ToMnet 也能通過經典的 Sally-Anne 測試，也就是說 ToMnet 可以意識到他人所擁有的錯誤信念(false belief)。在上述的實驗中，研究人員設計新的角色「近視者」放入虛擬空間中，而在測驗進行途中當施測者調整風景使近視者的視野變模糊時，ToMnet 可以對近視者接下來的行為做出準確的預測—近視者會依賴走過的路，而非像視力較好的角色策略性地得分。

當人工智慧理解人類 究竟是福是禍

人工智慧其實在「情緒辨識」方面已經有了不少成就，例如：許多臉部辨識軟體已經能夠偵測出人類極細微的臉部表情，分辨表情是否誠懇，甚至還能讀出一些像是性向、政治立場、智商等等的私人資訊；也有聲音辨識軟體可以從音調及語氣推測人類當下的情緒，觀察的細膩程度甚至遠超過人類對人類的了解。另外，人工智慧不像人類的記憶力及統計分析能力有限，它可以運用大數據整理一個人的朋友組成、瀏覽紀錄、聊天紀錄、情緒變化等等，一段時間過後 AI 便對那個人的興趣及習慣瞭若指掌，甚至可能比他自己都還要更了解他。

當 AI 理解人類，最直接的受惠者當然是人類。許多娛樂媒體企業(例如 Facebook、Youtube……)投資大量的工程師，想盡辦法分析大眾的喜好，以推出

更受歡迎的操作介面或作品；麻省理工旗下的媒體實驗室 Affectiva 嘗試在車廂內裝置 AI 情緒辨識系統，可以用紅外線和 RGB 相機追蹤駕駛者臉部和頭部位置，監督其是否精神不濟或漫不經心，也可以偵測駕駛者的情緒，若是過度興奮或生氣激動，AI 會判定並介入駕駛功能，替未來駕駛人增添了不少安全保障。

然而，當人工智慧的心智被開發後，伴隨的諸多威脅也令人堪憂。偏好時常影響決定，若 AI 理解人類，便懂得如何操控人們的思維，導致人們的判斷被偏好蒙蔽，更可怕的是，人們會因為 AI 提供的是自己喜歡的而依賴它，這也就是為什麼現今有這麼多手機成癮症患者的出現；另外，當 AI 能熟練地掌握人類的錯誤信念，便懂得如何輕鬆誘導人類掉進圈套之中，詐騙的問題自然將會如雨後春筍般湧現。

「水能載舟，亦能覆舟」AI 的進步就像一把雙面刃，但其實任何工具的好壞皆取決於使用者，就像世界頂級駭客米特尼克曾說過：「我對人工智慧不會有具體的立場，因為它本身是中立的，它最後到底是好是壞，取決於人類對它的應用。」再優秀的人工智慧一旦運用失當，也會變成用來作惡的利器。

參考資料：

1. Matthew Hutson, “Artificial intelligence has learned to probe the minds of other computers”, Science, 27 Jul 2018

2. Pete Etchells, “The Sally Anne task: a psychological experiment for a post-truth era”, The Guardian, 23 Jan 2017

3. Mikko Alasaarela, “The Rise of Emotionally Intelligent AI”, Median, 8 Oct 2017