擘劃一個以社會科學為核心導向的人工智慧研究願景

江彥生*

在過去二十年間，計算社會科學（Computational Social Science, CSS）已蔚為全球學術界的重要領域，[1] 它結合了社會科學的問題意識與計算方法的技術，讓研究者能處理過去無法觸及的龐大與複雜的有關個人與社會現象的資料。先不論歐美，就以我們身處的亞洲而言，已經有許多該領域的研究中心成立，包括最早成立的日本神戶大學的計算社會科學中心、新加坡國立大學的人文與計算社會科學中心、香港中文大學的計算社會科學研究群組，以及韓國高等科學技術研究院的數位人文與計算社會科學學程，此外在中國有更多相關的研究單位。相較於國際學術圈在 CSS 這塊領域上的快速發展，臺灣至今尚未有相關的研究中心，未來令人期待。

CSS 的研究在人工智慧（AI）時代更顯得重要。眾所皆知，AI 已經影響我們生活的各個層面，從自動駕駛、醫療診斷、金融決策、娛樂，乃至司法判決，AI 都已經深入這些決策過程。對臺灣學界而言，我們或許可以省思：在 AI 時代我們冀望何種 CSS 中心藍圖？而它們又如何有別於現有的典型 CSS 模式？

雖然計算方法與社會科學研究的結合已成為主流，但計算社會科學與社會計算（social computing）之間仍存在著些微差異。其分野在於，究竟主要核心在於回答實質的社會科學問題，抑或是僅專注於計算方法技術本身。在 AI 時代，這一差異更加關鍵，特別是臺灣當前規劃的方向，無論是經費或是教育投資，在後者的比重似乎遠超過前者。

在此脈絡下，我提出一個願景：建立一個以人文與社會科學為核心，來理解 AI、並引導未來 AI 設計的 CSS 研究中心。這一願景可從以下幾個方向加以闡述：

* 中央研究院社會學研究所研究員

[1] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915), 721-723.

一、用 AI 來拓展與深化社會科學的理論與模型

近年來，我們看到越來越多人工智慧（AI）模型，不僅強化，甚至顛覆了傳統社會科學的典範理論與模型。以下我提供幾個代表性的例子。

（一）案例一：民主 AI 與公共財的分配

英國的 DeepMind 公司（其創始者於 2024 年榮獲諾貝爾化學獎）旗下的資訊科學家與心理學家合作發表的 “Democratic AI” 一文，[2] 探討如何分配公共財的利益。「公共財」（public goods）是一個社會科學領域中的經典課題，其例子遍布於我們生活周遭：從社區的志工服務、國防安全的投入，到解決全球性氣候變遷的協議，都是公共財的議題。由於公共財的利益可以很多人共享，但是投入公共財需要成本，因此，從自利的角度來看，沒有誘因促使個人投身公益。社會科學數十年所累積的文獻指出，個人貢獻公共財的意願，取決於其手上擁有的資源多寡，而這也引發了關於富者是否應該按比例多付、以及如何在貢獻者之間回饋公共財利益的爭論。哲學家與實驗社會科學家提出了多種分配的理論和原則，每個理論都有不同的預測，在實驗研究中也獲得不同程度的實證支持。

DeepMind 團隊的作法巧妙地繞過這些理論上的爭辯，他們採取一種資料導向或者說是「民主導向」的策略：不去構想人們的偏好，而是直接讓他們表態。這些科學家首先邀請受試者參與一個公共財的遊戲實驗，藉此收集參與者的行為與偏好資料，接著再訓練一個 AI 來模擬人類在遊戲中的行為，藉此 AI 生成大量的行為資料，提供素材來訓練另一個核心 AI，讓它學習如何分配公共財的回饋，進而促使大家願意奉獻公共財。接著 DeepMind 團隊在後續的實驗中將這個 AI 模型與社會科學的典範模型相互比較，考驗不同的理論與 AI 模型所主導的分配方案，何者能受到人類受試者的青睞？測試結果顯示，AI 模型所推出的方案獲得大家支持的程度竟優於現有哲學或社會科學理論所推出的方案。換言之，在贏得人心支持上，該 AI 模型超越了既有的社會科學理論。

（二）案例二：用 AI 瓦解犯罪網絡

另一個例子來自網絡分析（network analysis）領域。舉例來說，在打擊組織犯罪時，警方常需要知道：如果要讓整個犯罪網絡瓦解，應該先抓誰？過去的做法，是參照網絡學者所設計的一些指標，例如看誰的關係最多（節點的連結程

[2] Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., ... & Summerfield, C. (2022). Human-centred mechanism design with Democratic AI. *Nature Human Behaviour*, 6(10), 1398-1407.

度高）、誰在不同群體之間位居中間人的角色（brokerage）、或是誰跟大家的網絡距離最短。無論成效如何，這些方法基本上都是學者專家「手工打造」（hand-crafted）的策略。

美國加州大學洛杉磯分校（UCLA）的資訊科學家團隊，嘗試讓 AI 自己學怎麼瓦解一個網絡。[3] 他們用一種叫做「強化學習」（reinforcement learning）的技術，結合圖神經網絡（graph neural network），讓 AI 在大量的模擬網絡資料中，藉由不斷的嘗試與修訂，學會找出該移除哪些節點（人物）能夠讓整個網絡的連結度很快下降，藉此瓦解集團成員之間的聯繫。

有趣的是，他們只用一種理論式的「無標度網絡」（scale-free network）模型 [4] 來訓練 AI，雖然只是用虛擬的網絡當作訓練教材，後來將訓練好的 AI 模型測試在瓦解真實的犯罪網絡上，表現得比專家學者所制定的策略還要好。這顯示 AI 不見得需要理解網絡理論，透過資料上的訓練依然能夠摸索出有效的瓦解策略。

（三）我的提議：從「黑箱」到新發現

這兩個案例都說明了：AI 在特定任務上（例如公共財分配、瓦解網絡）已經有能力超越人類專家。然而，這些 AI 的設計並不是基於社會科學理論，而是單純在訓練數據中追求最佳化的目標，至於它們「為什麼」會做出這些決策，往往是個黑箱。

要打開這個黑箱，一種方法是對 AI 本身做系統性的實驗。因為針對 AI 模型，我們可以用很低的成本快速在各種模擬情境下測試，讓我們能觀察它在不同刺激下會有何反應。累積這些經驗，或許能讓我們推敲出 AI 背後的運作原理，甚至從中發現新的社會現象，刺激我們發想新的社會理論。

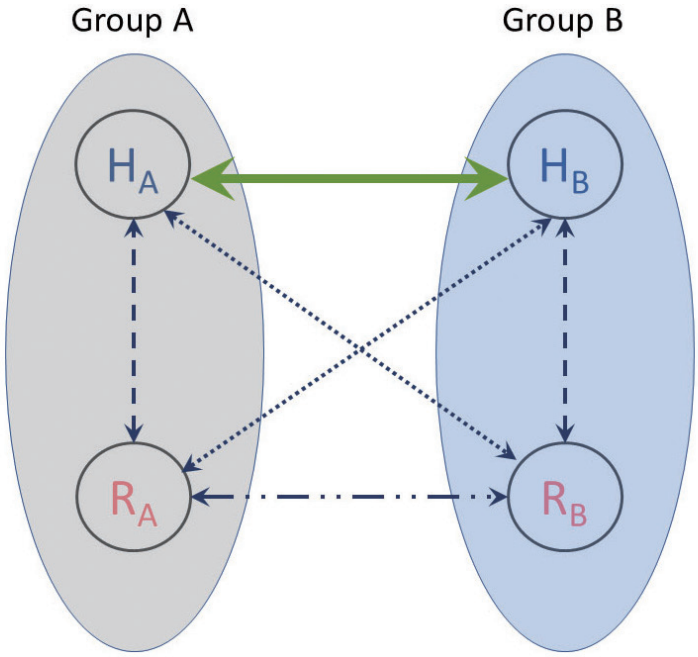
另一種相關的做法，是把 AI 放進真人參與的互動實驗裡。例如，我目前正在和同事合作，訓練一個強化學習式 AI，目標是減少兩位來自敵對陣營的人之間的敵意。我們的實驗設計中有四位參與者：兩位真人，以及兩個由 AI 控制的機器人，請參考圖一：

假設我們在實驗裡觀察到一個規律：當其中一個 AI（圖一中的 R_B）同時對兩位真人（H_A 和 H_B）表現出敵意時，這兩位本來彼此間有敵意反而變得更友好。

[3] Fan, C., Zeng, L., Sun, Y., & Liu, Y. Y. (2020). Finding key players in complex networks through deep reinforcement learning. *Nature Machine Intelligence*, 2(6), 317-324.

[4] Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939), 412-413.

雖然我們無法清楚得知 R_B「使用這招」其內部學習的邏輯是什麼，但它的行動所引發的人類反應，卻揭露了社會心理學中的一個經典理論：「敵人的敵人，就是朋友」。這就說明了將 AI 放進實驗中與真人互動，或許可以幫我們驗證甚至發現新的關於人際互動的社會科學理論。



圖一：規劃中的 AI 模型與實驗：有兩個敵對陣營（Group A 與 B），每個陣營中有兩位決策者，其中一位是真人，另一位是 AI。假設身處於一個社群媒體環境中，真人或是 AI 的身分無法辨別，因此決策者只知道自己和其他人是屬於哪個陣營，觀察他們的行為，但無法得知是否是真人或是 AI。

二、用社會科學的典範與方法來幫助我們理解 AI

近幾年，大型語言模型（Large Language Model, LLM）像是 GPT 系列的出現，可以說是一場劃時代的轉變。就像搜尋引擎改變了我們查找資訊的方式一樣，LLM 驅動的聊天機器人，也迅速成為我們工作與生活上的助手。由於這些 LLM 模型背後蘊含了數以億計的參數，要單純靠看它的內部結構來解釋它的行為，幾乎是不可能的事。因此，研究者開始改用「刺激—反應測試」的方法，透過精心設計的提問或情境，觀察 LLM 會怎麼回應，來推測它的行為模式。

（一）LLM 表現得像人類嗎？

一個重要的研究課題，是檢驗 LLM 是否會展現出人類在日常互動中常見的心理特徵，例如同理心、社會推理（猜測他人想法）、以及各種認知偏誤。在這方面，社會科學具備重要的貢獻：因為心理學、經濟學與社會學等領域早已發展出整套實驗設計與測量工具，用來研究人類行為，而這些方法完全可以套用到 LLM 身上。

（二）案例一：AI 具備人類的認知偏誤嗎？

德國認知科學家 Eric Schulz 與同事早期做過一個引人注目的研究，[5] 他們使用幾個經典的心理測驗來考驗 GPT-3：

- **Linda 謬誤：**測試人們在「連詞謬誤」（conjunction fallacy）上的偏誤。也就是，人們往往會覺得兩件事情同時發生的可能性，比單一事情發生的可能性還大，即使從數學機率上來說是不成立的。例如說，在實驗中給你一段關於名叫 Linda 的女性的描述：她大學時主修哲學、關心社會正義、參與反戰運動。接著問：「哪一種情況比較可能？」
  - ○ Linda 是銀行出納員
  - ○ Linda 是銀行出納員，而且是女權運動的活躍分子

  多數人會選第 2 項，因為它聽起來更符合描述，但其實「同時是出納員又是女權運動者」的情況，機率一定比「只是出納員」更小。這就是連詞謬誤。

- **Wason 卡片選擇任務：**用來檢驗資訊搜尋的策略。參與者會看到幾張印有字母與數字的卡片，並被要求檢視一個規則（例如說「如果卡片一面是母音，另一面必須是偶數」）。受試者被寄予的任務是翻最少數量的卡片來檢查規則是否成立。實驗發現人類常常選錯卡片，顯示我們在邏輯推理上的直覺並不總是正確。

- **多臂吃角子老虎（multi-armed bandit）遊戲：**想像你在賭場面前有好幾臺吃角子老虎機，每臺機器的中獎機率不同。你每次可以選一臺玩，繼續用中獎率不錯的那臺（exploitation），還是去試試看其他可能更好的機

[5] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.

臺（exploration）？這個測試是用來研究人在不確定環境下，在「探索新物」與「利用既有」之間的取捨。

實驗結果發現，GPT-3 在以上這些測驗上的表現，不僅接近人類水準，有些情況下甚至比人類更好。這讓人忍不住想問：AI 會不會比人類更「理性」？

（三）案例二：經濟學家理解 AI 的理性

經濟學家也做了類似的檢驗。[6] 他們用 GPT-3.5-turbo 參加一系列行為經濟學實驗，檢驗它是否符合個體經濟學中的核心理論，例如「顯示性偏好」（revealed preferences）與「效用最大化」。結果顯示，GPT 的決策，比起人類參與者，更符合經濟理性。

這又引出另一個問題：如果 LLM AI 的確比人類理性，那它是不是只擁有「人性中擅長優化的那一面」，而缺少人類那些非理性的一面；像是情感衝動、偏見、熱情，或是利他心？

（四）案例三：檢驗 AI 的「人格」？

為了檢驗 AI 是否具有人類非理性的特質，例如個性特質，學者讓 GPT-3 和 GPT-4 做了心理學界常用的 OCEAN「大五人格」測驗。[7] 結果顯示：

- **外向性**（Extraversion）：與人類差不多。
- **神經質**（Neuroticism）、**親和性**（Agreeableness）、**開放性**（Openness）：都比人類低。
- **盡責性**（Conscientiousness）：GPT-4 較高，GPT-3 較低。

（五）案例四：利他行為的實驗

另一組心理學家用知名的「獨裁者遊戲」（Dictator Game）測試 OpenAI 的 text-davinci-003 模型。[8] 在這個遊戲中，給予者（獨裁者）可以決定要分多少資源給受贈者，而後者不具備任何反駁或回饋的權利。在實驗中，AI 的受惠對象有時是人類，有時是另一個 AI。結果發現，這個 AI 比大多數人類更具利他心，而且有趣的是，它對另一個 AI 的慷慨程度，反而比對人類還要高。

---

[6] Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51), e2316205120.

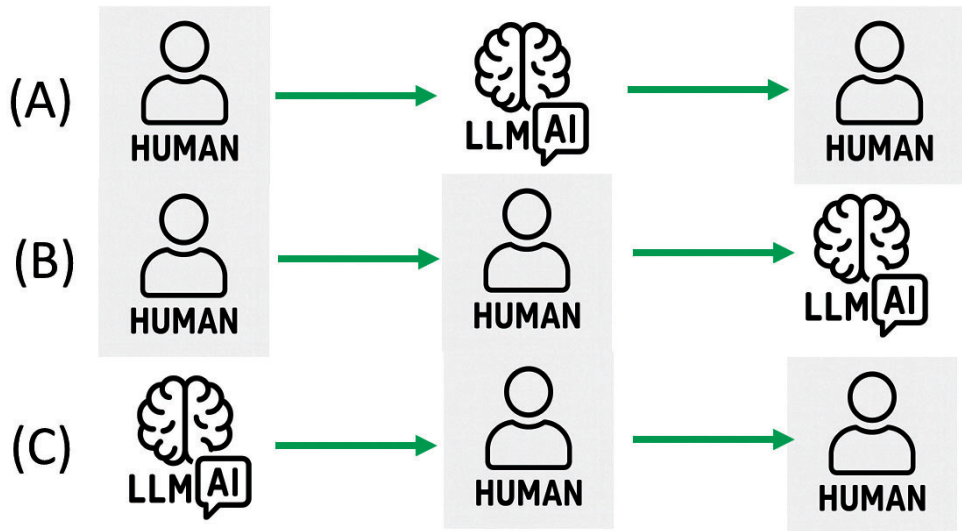
[7] Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), e2313925121.

[8] Johnson, T., & Obradovich, N. (2025). Testing for completions that simulate altruism in early language models. *Nature Human Behaviour*, https://doi.org/10.1038/s41562-025-02258-7

以上所有這類研究都有一個共同的發現：LLM 的表現對「提示詞」的設計非常敏感。指令只要改動，就有可能導致不同的結果。這讓「提示工程」（prompt engineering）成為未來的重要研究領域。

三、未來願景：從單一 AI 到「人機生態系」

目前的研究，大多還是集中在一個 AI 單獨的「人性化」表現。但我認為，下一步值得探索的，是把 AI 放進「人與 AI 混合生態」的脈絡中來觀察。畢竟，人類對 AI 可能還是存在不信任或偏見，而 AI 在團隊中的位置與角色，可能會直接影響決策結果。例如，在法律判決裡，我們應該先由 AI 提建議、再由人類做最後決策？還是反過來？或者以圖二為例，假設在團隊中有多位決策者，我們應該把 AI 放在核心主導位置，像是（A）或是（B），還是放在起始輔助角色，像是（C）？



圖二：人機混雜的三種決策模式：箭號代表傳遞判斷意見的方向。舉例來說，在（A）模式中，第一位人類做了一個初始的判斷，傳遞給中間的 AI，接著這個 AI 消化過後，再傳遞它的判斷意見給最後一位人類決策者做決定。

我認為以上問題的答案取決於兩件事：

1. **能力：**在特定領域，人類專家與 AI 誰的判斷更準確或效率更高？
2. **態度：**人類對 AI 的建議，是否信任並採納？

如果人類對 AI 的不信任很高，即使 AI 的答案很準確，也可能被忽視或扭曲。因此，要打造一個高效、和諧的人機生態，我們必須同時理解 AI 的決策能力，以及人類對它的心理反應。而關於這些問題，社會科學能在這課題上提供一些線索。畢竟，沒有其他領域比社會科學更了解人類在互動中，無論對象是另一個人，還是一臺機器，會怎麼想、怎麼做。