《科際整合月刊》，2026 年 9 月，11(9)，1-25。
Interdisciplinary Research Monthly. September 2026

人工智慧對社會科學研究的機會與挑戰

楊惟任*

摘要

人工智慧（尤其是大型語言模型與生成式 AI）已由研究輔助工具轉化為社會科學知識生產過程的方法論變項。本文以批判性文獻檢視為主體，輔以質性個案分析，聚焦三個研究問題：其一，大型語言模型與生成式 AI 究竟改變社會科學研究的哪些環節；其二，其偏差複製、模擬失真與黑箱性質如何影響研究效度與知識正當性；其三，何種治理機制能於研究過程被具體操作。本文研究結果指出，AI 的貢獻包括提高文本與非結構化資料的處理能量、支援前導性虛擬實驗與合成樣本、降低資料標註與文獻整理負擔，其限制涵蓋樣本代表性偏差、提示詞敏感度高、輸出具備非確定性，以及潛在的隱私風險與低可重製性。本文提出以透明度、公平性、可重製性與問責為核心的負責任 AI 應用框架，並將其轉譯為資料表、模型卡、分組評估、提示登錄與隱私衝擊評估等可檢核程序。本文的貢獻係將分散於 AI 倫理與計算社會科學文獻的原則性討論，落實為社會科學研究可操作的分析架構，以避免研究者過度追求技術效率，而忽視研究方法的嚴謹度。

關鍵詞：人工智慧、社會科學、大型語言模型、計算社會科學、負責任 AI

* 楊惟任教授，英國 University of Warwick 政治學暨國際關係博士，現任職於世新大學行政管理學系，曾任世新大學國際事務長、公共事務長、共同課程委員會主任委員，研究專長包括人工智慧與社會科學、川普學、全球治理、國際氣候政治。Email: william@mail.shu.edu.tw。
本論文經兩位雙向匿名審查通過。收件：2026/7/28。同意刊登：2026/8/16。

壹、前言

當代社會科學正經歷知識論與方法論的顛覆性變革。長期以來，社會科學仰賴質性詮釋、量化調查與田野觀察等經典方法，試圖解構人類行為、社會結構與制度變遷。然而，隨著機器學習、大型語言模型（Large Language Models, LLMs）及計算社會科學（Computational Social Science, CSS）的崛起，人工智慧（Artificial Intelligence, AI）已不再僅是輔助工具，而是逐步滲透至研究的整個過程－從資料探勘、假設生成到因果驗證。這場變革遠超越工具理性的層次，它代表著研究範式的根本性改變，並帶來方法論、倫理規範與社會實踐的多層次衝擊（Sarker and Furht, 2025）。

Microsoft（2023）的研究指出，AI 不僅加速知識的合成與實驗自動化，更成為跨領域合作的催化劑，重構科學探究的本質。早在 2009 年，Lazer 等人（2009）便提出 CSS 的理論框架，倡議利用數位足跡（digital footprints）研究社會互動。近年來，LLMs 的突破性發展則強化此趨勢，如 Argyle 等人（2023）運用 LLMs 模擬人類樣本進行政治傾向分析，展現 AI 在生成合成資料（synthetic data）的巨大潛力。值得一提的是，Lazer 等人（2020）於十年後重新檢視計算社會科學的進展，指出資料近用權限、隱私保護規範與跨機構協作機制等結構性障礙，仍是此一取徑能否累積知識的關鍵條件。此一提醒適用於當前以 LLMs 為核心的新一波研究方法變革。

放眼全球學術版圖，AI 雖已滲透社會科學諸多子領域，其應用樣態卻呈現不均衡的發展。教育學界利用 AI 打造個人化學習，卻憂心批判性思維的流失；經濟學與社會學則引入 AI 評估勞動市場變遷與自動化衝擊；數位人文領域利用其處理非結構化資料的能力，揭示傳統方法難以捕捉的宏觀社會模式，但這些進展背後，潛藏著不容忽視的結構性隱憂（Gefen, 2020）。

相較於早期將 AI 定位為提升效率的工具，當前研究聚焦於其背後的權力效應與對知識生產造成的影響。Chakravorti 等人（2025）的研究顯示，社會科學家雖普遍肯定 AI 在文獻整理、初步編碼與寫作輔助的效率，卻也擔憂缺乏透明度與過度依賴系統導致的認知偏誤，以及可能引發的研究能力退化與社會偏見深化。Bender 等人（2021）與 Blodgett 等人（2020）指出，大規模語言模型的數據資源、樣本偏誤、技術實踐或系統導入皆內嵌社會權力關係，若研究者未加辨識，模型輸出很容易再製與放大既有的性別、種族、階級與語言不平等。

因此，本文將社會科學面臨的困境界定為「規模」與「效度」之間的拉鋸。傳統調查、訪談與田野方法擅長深度描述，卻常受到樣本規模、時間成本與跨地域比較所限，LLMs 能快速生成假設、輔助文本編碼、模擬政策情境，並形成前導性的合成樣本。不過，Argyle 等人（2023）與 Aher 等人（2023）指出，模型即使能在特定情境重現群體層次的反應分布，也不能直接替代由親身經驗、制度位置與互動情境建構的人類行為。換言之，AI 在方法論的定位，應視為研究的輔助工具與初期探索，而非直接取代人類研究，更不應作為最終的論證依據。

本文不把 AI 簡化為機會與挑戰的二元對立，旨在將 AI 提升至制度性技術的層次，藉此重新審視其對研究設計、知識權威及倫理責任的根本性影響。本文的研究目的包括辨識 LLMs 與生成式 AI 介入社會科學研究的具體環節；說明偏差放大、模擬失真、隱私風險與可重製性困境如何轉化為方法效度問題；提出可檢核的負責任 AI 應用框架，以在技術效率與人文社會科學核心價值之間，建構兼顧兩者的平衡機制。

基此，本文將研究問題聚焦為三項相互扣連的問題：其一，在資料蒐集、標註、假設生成、前導實驗與研究報告等環節，LLMs 與生成式 AI 如何改變社會科學研究？其效率增益是否產生效度代價？其二，模型偏差、提示敏感度、非確定性輸出與模擬失真，如何威脅研究結論的內在效度、外在效度與規範正當性？其三，面對以上矛盾，哪些透明、公平、可重製且可問責的操作機制，能在研究過程落實負責任 AI 應用？

本文屬於批判性整合型論文，目的不是提出全新的經驗理論，亦非進行完整的系統性回顧，其預期貢獻在於比較、批判並重組既有文獻中分散出現的「典範轉移」、「演算法偏差」、「合成樣本」、「研究流程自動化」與「AI 倫理治理」等論點，並將抽象倫理原則轉化為社會科學研究情境的具體檢核機制。相較於僅主張 AI 可提升效率或帶來風險的論述，本文著重回答：「當研究者將 AI 納入研究設計時，哪些程序必須被記錄、驗證與問責」。

本文的理論架構為「雙重辯證」（dual dialectic）。第一層是方法論辯證，係「規模與效率」和「效度與詮釋深度」之間的角力：AI 大幅提高資料處理與模擬能力，卻可能犧牲情境詮釋、因果識別與推論穩健性。第二層是規範辯證，即「技術加速主義」與「人文社會科學核心價值」之間的衝突：自動化與生產力誘因必須與公平、

隱私、透明、問責及學術誠信一併檢視。後續論述，包含文獻探討、機會分析、風險辨識與框架建構，皆依循此一架構展開。

本文採用批判性文獻檢視結合質性個案分析的研究設計。文獻評述除了進行主題歸類，亦比較既有研究如何界定 AI 在社會科學研究的方法角色、效度風險與倫理責任，以辨識其論證盲點與可整合之處。資料來源以同儕審查期刊論文、頂尖會議論文、可查證工作論文及具公信力之政策文件為主，並排除來源不明、出版資訊不完整或與本文論點關聯薄弱的資料。文獻蒐集與篩選程序參酌「系統性回顧與統合分析優先報告項目 2020」（簡稱「PRISMA 2020 聲明」）（Page et al., 2021）之精神，惟未涉入效果量之統合分析或完整之系統性文獻回顧。檢索詞涵蓋「人工智慧」、「大型語言模型」、「生成式 AI」、「社會科學」、「計算社會科學」、「研究方法」與「研究倫理」等中英文關鍵詞。參考文獻以 2020 年至 2026 年間發表者為主，並酌予保留計算社會科學與 AI 倫理領域的基礎文獻，其內容須明確涉及 AI 對研究方法、資料處理、模擬實驗或倫理治理的影響，且具備可查證的出版資訊或穩定連結。

在檢索範圍上，本文以 Web of Science、Scopus 與 Google Scholar 作為主要英文文獻來源，並輔以 ACL Anthology、arXiv 及各出版社資料庫，中文文獻則檢索華藝線上圖書館與國家圖書館「臺灣期刊論文索引系統」。前述排除標準的具體判準為：「來源不明」指無法辨識作者或出版機構者，「出版資訊不完整」指欠缺卷期、頁碼或穩定識別碼（如 DOI）者，「關聯薄弱」指未實質涉及 AI 對研究方法、資料處理、模擬實驗或倫理治理之討論者。經去除重複、逐筆檢視全文並剔除不符標準者後，最終納入深入分析的核心文獻約 50 篇，並依「方法增能」、「模擬實驗」與「規範治理」進行歸類。

個案選擇採理論性抽樣邏輯，其目的並非追求統計代表性，而是呈現 AI 介入社會科學研究之主要方法類型。本文選取 Argyle 等人（2023）之研究與史丹佛大學模擬人類受試者項目，分別代表「以條件化提示重建群體回應分布」與「以大規模虛擬代理預測干預效果」兩種方法取向。個案分析依循四個維度展開：方法設計與條件式提示策略、效度宣稱及其驗證方式、偏差與外推風險，以及對研究設計與倫理治理之啟示。整體分析策略首先歸納 AI 介入研究設計之主要路徑，接著透過個案檢視其方法增益與效度限制，最後將倫理原則轉譯為可觀察之指標。本文並不以實驗資料檢定框架成效，亦不主張提出之指標足以一次性解決所有 AI 治理問題，而是為後續經驗科學研究提供一個可供驗證、修正與擴充的分析架構。

貳、文獻探討

本節並不單純羅列文獻，而是將近年 AI 與社會科學研究的討論歸納為三條路徑：第一，將 LLMs 作為資料標註、文本分類與跨語言分析工具的「方法增能」路徑；第二，將 LLMs 作為合成樣本、虛擬代理與前導實驗的「模擬實驗」路徑；第三，將 LLMs 視為可能重製不平等、降低可重製性並改變責任歸屬的「規範治理」路徑，這三條路徑共同構成本文後續分析的文獻基礎。

相較於單純整理應用案例，本文更著重既有研究之間的分歧：方法增能研究強調效率與規模，模擬實驗研究凸顯前導測試與成本降低，但倫理治理文獻顯示，若缺乏明確概念、外部驗證與可追溯紀錄，效率優勢可能轉化為方法風險。這個矛盾正是本文提出雙重辯證與負責任 AI 框架的立論基礎。

一、AI 在社會科學研究的機會與創新應用

在方法增能路徑，LLMs 展現改造計算社會科學流程的能力。Ziems 等人（2024）以 13 種語言模型與 25 項 CSS 任務進行評估，指出 LLMs 可在零樣本或少樣本情境下支援文獻分類、社會概念標註與解釋生成，但其表現並非在所有任務皆優於微調模型。這意味著 LLMs 的應用範疇並非「自動取代人工編碼」，而是作為人類標註、理論判讀與錯誤檢核的協作工具。

Thapa 等人（2025）指出，LLMs 可應用於情緒分析、仇恨言論偵測、事件理解與社會網絡分析等任務，協助研究者處理社群媒體與政策文件等非結構化資料。然而，這類應用必須區分「分類準確率」與「社會概念測量效度」，模型能準確貼上標籤，並不必然表示研究者所欲測量的概念已被妥善操作化，此區分使 LLMs 的使用回到社會科學方法論的關鍵問題，即概念化、測量與推論之間的關係。

Microsoft（2025）在《社會責任人工智慧：研究的挑戰與機遇》報告，闡述 AI 對方法論的轉型潛力，並引用 Argyle 等人（2023）的模擬研究，展示 AI 在政治分析與社會行為預測的應用價值。更有甚者，生成式 AI 可動態產製合成資料集，用於測試社會理論或模擬群體動態，為「虛擬實驗室」的概念奠定基礎。經濟合作暨發展組織（Organization for Economic Co-operation and Development, OECD）亦指出，AI 在科學生產力的潛力，特別適用於社會科學的資料密集型任務，如跨國政策影響評估或系統性不平等的分析（OECD, 2024）。

個案研究佐證這些潛力。Argyle 等人（2023）在 *Political Analysis* 提出的創新方法顯示，透過精確的條件提示（conditioning prompt），LLMs 可充當特定人口子群體的模擬受試者，模擬不同群體對政策議題的回應分佈，該研究利用 GPT-3 生成與真實調查高度相符的合成樣本，揭示 AI 在降低研究成本同時維持代表性的可能性。

史丹佛大學的研究團隊則將此方向推進得更遠，他們利用 GPT-4 模擬美國代表性樣本對 476 項隨機化干預的反應，結果顯示 LLMs 的預測準確度與人類專家相當，且與實際干預效果的相關性高達 0.85。更引人注目的是，該團隊建構包含 1,052 名虛擬個體的代理系統，能準確再現真實調查回應達 85%，這一成果意味著虛擬實驗環境的可行性，特別適用於那些成本高昂或具倫理敏感性的研究設計（Ashokkumar et al., 2026; Park et al., 2024）。

二、方法論與實務挑戰

儘管工具潛力廣受期待，AI 在社會科學的應用首先面臨資料品質、資料污染與代表性不足的挑戰。Bender 等人（2021）聲稱，LLMs 常以規模擴張掩蓋數據來源、資料清理與環境成本問題，Blodgett 等人（2020）警告，偏差研究若未釐清受害群體、危害類型與規範基礎，容易以技術指標取代社會診斷，Kalhor 等人（2025）強調，模型在重構學術關係時，容易衍生族群與性別的結構性偏誤。這些研究皆說明：AI 的輸出不是中性資料，而是由訓練數據、標註實務與社會權力構成的再現結果。

其次，研究者專業技能與思維能力的潛在退化，亦構成 AI 應用不容忽視的風險。Microsoft（2025）指出，研究工作若過度依賴 AI 所提供的自動化分析與生成結果，可能降低研究者對方法論細節、資料處理流程與分析假設的掌握，因而加劇自動化偏差（automation bias）。此外，生成式 AI 所產出的內容在真實性與可驗證性上仍存在顯著限制，其經常生成表面上合乎邏輯、實則缺乏實證支撐的「幻覺」（hallucinations），不僅誤導研究判斷，亦對學術誠信與研究品質構成直接威脅。

再次，個案研究凸顯這些挑戰的現實性。Argyle 等人（2023）的模擬雖具創新性，但條件化過程的調整需要深厚的技術與理論深度，否則極易導致偏差放大。史丹佛大學的案例則揭示另一層問題，即 LLMs 傾向產生「安全」或平均化的回應，表現出明顯的順從性偏差，因此研究團隊建議採取「人類＋LLMs」並行的混合設計，以平衡 AI 效率與人類判斷的必要性。

三、倫理、偏差與隱私議題

倫理議題已然躍升為當代 AI 治理的重要關切。Chakravorti 等人（2025）以 284 份問卷與 15 位社會科學家的半結構式訪談為基礎，指出生成式 AI 在研究情境的主要風險包括自動化偏差、研究者去技能化、研究不當行為、黑箱解釋困難與代表性傷害。這個說法呼應 Blodgett 等人（2020）對自然語言處理（Natural Language Processing, NLP）偏差研究的批判，偏差並非可由單一技術修補的缺陷，而必須將其放到受影響群體、制度脈絡與知識權力的架構中進行理解。

Chen（2020）指出，AI 不是中立工具，訓練資料嵌入的社會偏差在輸出中獲得科學化的合法性偽裝，進而強化刻板印象與邊緣化效應。聯合國教科文組織（United Nations Educational, Scientific and Cultural Organization, UNESCO）的實證研究顯示，當代 LLMs 持續產出涉及性別、性傾向與種族的系統性偏見，某些模型對女性職業選擇、LGBTQ+ 認同與少數族裔的呈現，存在明顯的負面傾向（UNESCO, 2024）。

隱私風險與資料安全也是重大課題。模型訓練仰賴龐大的個人資料，使得隱私洩露與資料安全成為 AI 應用的首要挑戰之一，敏感個人資訊可能被模型「記憶」，並在生成文本中無意間洩露（Ivanov, 2025）。此外，Microsoft（2025）警告 LLMs 可能產生誤導性資訊，甚至出現「價值對齊偽裝」（alignment faking），模型看似認同某種倫理立場，實則隱藏著相反的邏輯傾向。

Sarker 與 Furht（2025）提出整合透明度、公平性與隱私的倫理框架，經濟合作暨發展組織（Organization for Economic Co-operation and Development, OECD）的《人工智慧原則》（*OECD AI Principles*）、聯合國教科文組織（United Nations Educational, Scientific and Cultural Organization, UNESCO）《人工智慧倫理建議書》（*Recommendation on the Ethics of AI*）、Microsoft《負責任 AI 標準》（*Responsible AI Standard*）均倡議跨域合作與風險治理。反之，這些框架多屬一般性原則，若未轉化為社會科學研究的資料、模型、提示、驗證與報告程序，仍難以回應期刊審查、同儕複核與研究倫理審查的具體需求。

綜上所述，既有文獻已充分揭示 AI 在效率提升、資料標註、模擬實驗與治理實踐的潛力與風險，但相關研究鮮少回答三個值得注意的問題：AI 究竟改變研究設計的哪些節點？這些改變如何影響社會科學的效度裁量標準？研究者如何透過可稽核

程序管理相關風險？以下，本文將以此缺口為主軸，釐清 AI 的方法論革新，接著辨識其效度與倫理限制，然後提出可操作化的分析框架。

參、人工智慧對社會科學研究的機會

本節聚焦探討 AI，特別是 LLMs 與生成式 AI 對社會科學研究提供的機會，說明 AI 如何提升研究效率、擴大資料規模、促進方法創新及強化包容性。內容分為四個主要面向：（1）資料分析與處理能力的提升；（2）模擬與合成資料生成的創新應用；（3）研究流程自動化與生產力增強；（4）促進跨領域與包容性研究。透過代表性個案，說明這些機會的實務價值，並連結文獻探討的基礎。

一、資料分析與處理能力的提升

傳統社會科學研究長期受限於人工編碼的時間與成本瓶頸，AI 的應用使資料規模、資料類型與分析速度同時擴張。LLMs 可支援情緒分析、議題判讀、立場辨識與初步質性編碼，並從社群媒體、訪談紀錄、政策文本與歷史檔案中萃取可供後續理論分析的訊號（Thapa et al., 2025），但此處所謂「訊號」須經由研究者的概念界定與效度檢驗，不能直接等同於社會事實。

更重要的是，LLMs 對上下文與語用關係具有一定程度的掌握，使其相較於傳統詞彙模型或淺層分類器，更能處理諷刺、立場轉換、敘事框架與跨語言表述等複雜文獻現象（Ziems et al., 2024）。然而，跨語言能力不應被理解為文化等價性的保證，研究者仍需建立人工複核樣本、交互編碼信度與語境敏感的錯誤分析，否則容易將語言轉換誤認為概念測量。

二、模擬與合成資料生成的創新應用

透過條件提示工程（prompt engineering），LLMs 可作為特定社會群體的代理以生成合成樣本，形成「虛擬實驗」或前導測試的可能性。Argyle 等人（2023）提出的合成樣本、Aher 等人（2023）提出的圖靈實驗（Turing Experiments），以及 Horton（2023）以 LLMs 建構模擬經濟行動者的嘗試，均顯示 AI 已不只是文本處理工具，而逐步進入研究設計與假設探索階段。以下以 Argyle 等人（2023）的研究與史丹佛大學模擬人類受試者項目作為分析個案。

個案一：Argyle 等人的研究

Argyle 等人（2023）的研究提出，透過精心設計的條件提示，GPT-3 可模擬真實調查中不同人口統計群體的回應分佈。該研究的革新即是，它不僅降低傳統抽樣的成本，更解決代表性不足的難題，生成的合成樣本能用於初步假設測試、政策模擬與行為預測，這大幅降低資源受限的研究者的准入門檻，使其得以涉足具規模且高成本的研究領域。不過，所生成樣本的品質與模型版本、提示策略密切相關，需要高度謹慎的驗證。

個案二：史丹佛大學模擬人類受試者項目

史丹佛大學團隊以 GPT-4 模擬美國代表性樣本對 476 項隨機化干預的反應，並以既有實驗結果與人類專家預測作為比較基準。研究結果顯示，LLMs 具備前導預測的潛力，但人類資料仍不可或缺。此一突破並非意味 AI 可取代人類受試者，而是提供低成本的前測機制，使研究者能先以 AI 估計樣本規模、預測干預因子或檢驗題項清晰度，再進入正式的人類受試者研究。

若依前述分析維度比較前兩個個案，可看出其方法論意涵的層次差異。就方法設計而言，Argyle 等人（2023）透過細緻的條件化提示重建群體分布，史丹佛團隊則將單點模擬擴展為大規模干預預測；就效度而言，前者主要比較合成樣本與真實調查分布的吻合程度，後者則以既有實驗效果與專家預測作為外部基準。然而，兩者皆僅能證實群體分布或平均效果的相似性，無法等同於個體行為、制度脈絡與互動過程的完全一致。因此，本文將 AI 模擬定位為假設生成、前導估計與研究設計壓力測試，而非可獨立取代人類樣本的資料來源。

三、研究流程自動化與生產力增強

AI 的另一項貢獻是協助研究者處理重複性、探勘性與文件化工作，例如文獻追蹤、程式除錯、初步分類、摘要比對與方法附錄草擬。Bail（2024）指出，生成式 AI 有助於改善調查研究、線上實驗、自動內容分析與代理模型，Kusumegi 等人（2025）以大規模資料說明，LLMs 的使用已改變科學生產流程，但此類生產力提升只有在研究者保留概念判斷、資料檢核與因果推論責任時，才具有學術價值，若研究者放棄對資料處理流程與模型假設的掌握，自動化將轉化為方法風險。

四、促進跨領域與包容性研究

AI 亦具有促進跨學科整合的潛力，當研究問題涉及氣候變遷、不平等、數位治理或公共健康等複合議題時，LLMs 可協助整合政策文件、媒體資料、開放資料與訪談內容，並在初步階段生成可供研究者檢驗的概念圖像（Bail, 2024）。儘管如此，跨域整合並不意味理論邊界可以被抹除，研究者仍須說明資料來源之間的可比性、概念轉譯的前提，以及不同學科證據標準的差異。

此特性對於研究邊陲或資料匱乏的群體具有方法啟發，但其倫理前提必須更嚴格。以模型「代言」邊緣群體，容易將少數者經驗壓縮為數據的平均化印象，甚至以看似多元的輸出掩蓋樣本偏誤。故 AI 在包容性研究的合理角色，應是協助研究者辨識資料缺口、形成訪談問題與設計樣本補強策略，而不是替代研究者與受影響群體的實際互動。

表 1

AI 導入社會科學研究的方法與效度

研究執行步驟	AI 協同作業模式	方法論效能提升	查核事項
問題界定與假設生成	文獻摘要、概念對照、控制條件變異模擬	快速形成可檢驗命題	由研究者審核概念定義、理論前提與研究邊界
資料蒐集與標註	自動編碼、情緒/立場/主題分類、跨語言初譯	擴大非結構化資料處理規模	人工複核、交互編碼信度、錯誤類型分析與偏差檢測
前導實驗與模擬	合成樣本、虛擬代理與政策干預預測	降低前測成本並篩選研究設計	與人類樣本、歷史資料或既有實驗結果進行外部驗證
分析與詮釋	輔助質性編碼、提出替代性解釋、敏感度分析	優化數據探勘效益與證偽檢驗能力	記錄提示、模型版本與抽樣參數，保留反例與人類判斷紀錄
報告與再現	方法附錄草擬、提示登錄、資料表與模型卡	提升透明度與同儕複核可能性	公開可分享資料與程式碼，執行隱私衝擊與倫理審查

資料來源：作者自行整理

肆、人工智慧對社會科學研究的挑戰

本節聚焦探討 AI，特別是 LLMs 與生成式 AI 就社會科學研究帶來的挑戰，說明 AI 應用所帶來的風險與限制。內容分為四個主要面向：（1）偏差與公平性問題；（2）模擬人類行為的限制與可靠性挑戰；（3）倫理、隱私與問責議題；（4）方法論、透明度與可重製性困境。透過代表性個案，說明這些挑戰的表現，並對應前一節關於人工智慧對社會科學研究的機會的討論。

一、偏差與公平性問題

LLMs 在訓練過程呈現的偏差，不宜僅被理解為資料量不足或模型尚未最佳化的技術缺陷，而應視為既有社會結構在語言訓練樣本的再現。由於訓練資料高度集中於英語語境、高收入國家與主流文化文獻，模型處理低資源語言、非西方經驗或弱勢群體敘事時，經常出現覆蓋不足、刻板化或平均化輸出。Blodgett 等人（2020）提醒，若研究者未說明偏差的規範含義與受害對象，偏差測量可能淪為抽象指標，而非有效的社會結構問題的批判性分析。

Chen（2020）指出，AI 應用最隱蔽的風險乃它賦予社會偏差以「科學」的合法性，當模型輸出被認定為基於數據驅動的「客觀」結論時，潛藏的偏差便被隱沒在統計數字的背後。聯合國教科文組織的實證研究具體呈現這一現象：當代 LLMs 在性別、性傾向與種族議題持續產生系統性偏見，這些偏見被包裝成看似中性的文本（UNESCO, 2024）。

史丹佛大學的研究團隊亦指出，他們的模擬人類受試者模型存在明顯的順從性偏差，對特定社會群體的呈現遠不準確。Argyle 等人（2023）的條件化方法若未充分包容邊緣群體，亦容易強化既有不平等，這些限制不僅影響研究的內在效度，更可能誤導政策制定（Ashokkumar et al., 2026）。惟須補充，Park 等人（2024）發現，相較於僅以人口統計標籤條件化的代理，其以深度訪談為基礎的生成式代理架構可降低跨種族與意識形態群體的準確度偏差，這顯示，偏差程度取決於條件化資訊的品質與細緻度，並非模型不可避免的固定特性。

二、模擬人類行為的限制與可靠性挑戰

AI 模擬人類行為的根本限制，不外乎它不具備人類行為所依賴的具身性（embodiment）、制度位置、互動歷史與情感承擔。Bail（2024）指出，生成式 AI 可

改善若干社會科學方法，但其輸出不能替代因果識別、研究設計與外部驗證。Aher 等人（2023）也發現，LLMs 可重現若干具代表性的實驗研究結果，也可能在特定情境產生「過度準確」的扭曲，顯示模型對人類變異性的掌握仍然有限。

提示敏感度（prompt sensitivity）降低模擬結果的穩定性。細微的措辭、角色設定、溫度參數或模型版本差異，皆可能導致輸出顯著改變。Ziems 等人（2024）指出，LLMs 在某些分類任務可接近人類標註水準，但其錯誤模式與人類判斷並不相同，因此表面的準確率不能取代對錯誤分布、概念有效性與跨群體穩健性的檢驗。

就實證量級而言，Sclar 等人（2024）的系統性評估發現，僅是提示格式等保留語意的細微更動，即可能使模型表現產生高達數十個百分點的落差，而溫度參數越高，輸出的隨機性與離散程度越大，反覆抽樣所得的合成分布亦隨之漂移。依此，前述兩個個案的效度並非模型的固定屬性：Argyle 等人（2023）以 GPT-3 重建的群體回應分布，與史丹佛團隊改用 GPT-4 後的表現不宜等量齊觀：模型世代更迭雖普遍提升平均預測準確度，卻改變其偏差結構與拒答傾向，使得「合成樣本代表性」在本質上成為繫於特定模型版本、系統提示與抽樣設定的條件性宣稱，而非可跨版本穩定外推的性質。因此，任何以合成樣本為基礎的推論，皆應載明模型版本、生成日期與參數設定，並於版本更新後重行驗證。

個案研究深入揭示這些限制的現實性。史丹佛大學的虛擬代理在模擬人類樣本時展現出「安全回應傾向」，趨向於產生社會認可的答案，而非真實多樣的觀點，Argyle 等人（2023）的條件化方法即使創新，但條件化參數的微調差異會導致結果的實質變化。這表明 AI 模擬在當前發展階段難以完全取代人類樣本，而應定位為補充性方法。

此一動態性在量化社會科學的因果識別脈絡中尤其值得警惕。當研究者將 AI 生成的合成資料作為變項納入迴歸模型或實驗分析時，模型輸出本質上是對真實反應的「預測值」，而非直接觀測資料。若未經校正即逐行代入，等同於引入測量誤差，可能造成係數衰減、遺漏變數偏誤與內生性問題。Bisbee 等人（2024）以實證資料指出，LLMs 生成的合成樣本常系統性壓縮群體內變異，使有效樣本數低於名目規模，進而低估標準誤、高估統計顯著性，導致頻率式或貝氏式推論失真，若多個合成變項源自同一模型，彼此高度相關，也可能引發共線性並扭曲淨效果估計。

針對此類「生成迴歸子」問題，Angelopoulos 等人（2023）提出的預測賦能推論提供了一項可行方法：以少量真實標註資料校正模型預測所造成的偏誤，並在維持統計效度的前提下擴充樣本規模。因此，AI 合成資料在計量經濟學或實證社會學中的合理定位，不是直接取代觀測資料，而是在具備偏誤校正與外部驗證的研究設計中，作為輔助性的推論來源。

三、倫理、隱私與問責議題

LLMs 的訓練涉及龐大的個人資料，這就構成隱私的潛在威脅。Ivanov（2025）指出，資料洩露風險已成為 AI 應用的首要倫理隱憂，訓練資料中的敏感個人資訊可能被模型無意記憶，進而在生成過程被洩露。

更棘手的是問責與透明度的缺失。LLMs 的決策不透明，讓我們難以準確追蹤錯誤的來源或糾正偏差的方式。Chakravorti 等人（2025）表明，社會科學家對生成式 AI 的黑箱特性持普遍疑慮，這種不透明性不僅阻礙學術評審，更可能成為學術倫理瑕疵的溫床。

此外，AI 應用亦涉及智慧財產權爭議、錯誤資訊傳播、研究資料外洩，以及對弱勢群體的再次傷害。Bender 等人（2021）強調，大規模語言模型的環境成本、資料攫取與語言代表性問題，往往由最少受益的群體承擔最多風險。這類風險並非研究完成後才需處理的附帶事項，而應在研究問題形成、資料蒐集、模型選用與結果報告階段即被納入設計。

四、方法論、透明度與可重製性困境

LLMs 的非確定性輸出對傳統科學方法論構成直接挑戰。可重製性是科學知識累積的基礎，但當模型版本更新、提示措辭改變、系統提示未批露或抽樣參數不一致，先前結果可能難以重現。因此，Sarker 和 Furht（2025）主張 AI 研究報告必須提供比傳統研究更完整的方法資訊，包括模型名稱與版本、確切提示、抽樣參數、資料前處理程序、人工複核比例與錯誤分析策略。

然而，商業模型的專有性質讓完整透明度難以實現。研究者即使揭露提示與參數，也未必能得知訓練資料、過濾規則或模型更新細節。為降低這個限制，本文主張採取「最低可重製性揭露」，亦即保存提示與回應紀錄、標明模型版本與日期、提供人工驗證樣本、揭露失敗案例，並於可能情況下以開源模型或可公開資料進行穩

健性檢查。此一最低標準可使研究結果具有可追溯性，而不至於完全依賴不可審查的黑箱輸出。

伍、負責任 AI 應用框架的建構與操作

前述機會與挑戰若僅止於分類羅列，無法回應學術社群關心的「框架尚未操作化」問題。本文據此將負責任 AI 應用框架建構為四個可檢核的標的：透明度、公平性、可重製性與問責，其對應研究過程的資料蒐集、模型選用、分析執行、結果報告與後續複核，並將 Floridi 與 Cowls（2019）、Jobin 等人（2019）、OECD（2024）、UNESCO（2021）、Microsoft（2022）以及我國行政院（2023）揭示的倫理原則，落實為社會科學研究的實踐步驟。

第一項檢核標的為資料與模型透明度，其操作重點要求研究者在方法報告中檢具標準化文件。就資料而言，採用資料表（datasheets）揭露資料來源、組成、蒐集脈絡、標註流程與已知限制（Gebru et al., 2021），就模型而言，採用模型卡（model cards）載明模型版本、適用範圍、效能邊界與已測得偏差（Mitchell et al., 2019）。可觀察指標包括資料來源揭露完整度、模型版本與參數登錄率、提示語與程式碼公開比例，以及失敗案例是否被報告。

第二項檢核標的為偏差審查與公平性保障，其操作工具包括分組評估、對抗性測試、邊緣群體對照與錯誤類型分析。研究者應就性別、族裔、語言、年齡、社經地位與地區等面向檢視模型表現，並說明哪些群體可能因資料缺席或模型錯誤而受損。可觀察指標包括跨群體效能落差、低資源語言覆蓋率、偏差緩解措施之紀錄，以及受影響群體是否被納入評估過程。

第三項檢核標的為可重製性與問責。鑑於 LLMs 輸出的非確定性，本框架要求研究者完整記錄模型版本、系統日期、確切提示、抽樣參數、人工驗證流程與排除案例，並於可行範圍內公開資料與程式碼。可觀察指標包括可重製性查核表完成度、提示與回應紀錄保存率、隱私衝擊評估執行率，以及人類研究者於關鍵推論環節的介入紀錄。透過這樣的安排，AI 的使用責任不會被轉移給模型，而仍由研究者、研究機構與出版系統共同承擔。

第四項檢核標的為倫理審查與隱私保護。當研究使用受試者資料、社群媒體資料或具可識別風險的文本時，應先判斷資料是否可被合理再識別，並評估模型輸入是否可能造成資料外洩。操作工具包括資料最小化、去識別化、敏感資料排除、隱

私衝擊評估與 AI 使用揭露。可觀察指標包括敏感資料處理紀錄、研究倫理審查結果、資料保留與刪除規範，以及向受試者或資料提供者揭露 AI 使用方式的程度。

表 2

負責任 AI 應用框架的指標與操作對照

指標	對應倫理原則	操作工具	可觀察指標
透明度	透明、可解釋	資料表、模型卡、提示與版本登錄	資料來源揭露完整度、模型版本與參數登錄率、提示與程式碼公開比例、失敗案例報告
公平性	公平、不傷害	分組評估、對抗性測試、邊緣群體對照、錯誤類型分析	跨群體效能落差、低資源語言覆蓋率、偏差緩解紀錄、受影響群體納入程度
可重製性與問責	問責、學術誠信	可重製性查核表、提示/版本/參數登錄、資料與程式碼公開、人為介入紀錄	查核表完成度、提示紀錄保存率、可公開資料比例、關鍵推論之審核紀錄
倫理審查與隱私	隱私、知情、資料最小化	去識別化、敏感資料排除、隱私衝擊評估、AI 使用揭露	隱私衝擊評估執行率、敏感資料處理紀錄、倫理審查結果、受試者或資料提供者揭露程度

資料來源：作者自行整理

審查意見指出，前述四項檢核標的所援引的資料表、模型卡、分組評估與對抗性測試等工具，主要源自計算科學與演算法治理。若未經轉譯即套用於重視反思性、主體性與非結構化材料的質性研究，可能造成方法論斷裂。本文主張，該框架在質性研究中的作用，並非以標準化查核表取代詮釋，而是將量化評估中的透明度與問責精神，轉化為質性研究既有的嚴謹性判準。

具體而言，「資料表」可轉化為研究者對語料生成脈絡與自身定位的揭露，說明訪談情境、AI 介入環節，以及其可能造成的詮釋框架效應；「模型卡」可轉化為分析工具紀錄，載明模型版本、提示語，以及模型在編碼、摘要或主題萃取中的角色；「可重製性查核」則對應質性研究重視的稽核軌跡，要求保存原始語料、AI 輸出與研究者判讀之間的轉換過程，使詮釋鏈可供同儕檢視。至於「分組評估」與「錯誤分析」，在質性研究中應轉化為對沉默樣本、邊緣敘事與反例的系統追索，以避免模型的平均化輸出掩蓋少數者的異質經驗。

關鍵在於，這些工具應發揮輔助功能，而非取代研究者作出判斷。研究者必須保留概念界定、脈絡判讀與意義賦予的最終責任。當 AI 用於初步編碼或文本歸類時，質性研究者仍應透過人工複核、反覆閱讀與情境對照維持詮釋深度，並在方法論上區分「模型可辨識的表層語言特徵」與「研究者欲理解的深層社會意義」。如此，負責任 AI 框架才能在不削弱質性研究反身性的前提下，提升其透明度與可信度（Chakravorti et al., 2025; Davidson & Karell, 2025）。

為回應「框架僅止於規範性宣告」之疑慮，本文以表 2 的檢核標的，對本文選取的兩個個案進行後設應用回溯，示範該框架如何被用以診斷既有研究，而非僅作為未來研究的期許。

就個案一（Argyle 等人，2023）而言，在「透明度」標的上，該研究明確報告了模型版本（GPT-3）、條件化提示的建構邏輯與比較基準，於「提示與版本登錄」一項表現良好，惟其對逐題提示全文與完整程式碼的公開程度有限，且未附標準化的資料表與模型卡，故「提示與程式碼公開比例」僅屬中等。在「公平性」標的上，該研究以美國選民樣本為對象，未涵蓋低資源語言與非西方群體，「低資源語言覆蓋率」趨近於零，且缺乏系統性的分組效能落差分析，顯示其結論的外推須受嚴格限制。在「可重製性與問責」標的上，由於 GPT-3 已停止服務，其結果在後續模型世代難以完全重現，凸顯「最低可重製性揭露」的必要。

就個案二（史丹佛項目；Park 等人，2024；Ashokkumar 等人，2026）而言，其在「透明度」與「可重製性」標的上表現相對突出：研究團隊公開了代理架構、評估構念與部分程式碼，並以受限 API 保護受試者隱私，於「隱私衝擊評估」一項具參考價值。在「公平性」標的上，Park 等人（2024）主動報告並比較跨種族與意識形態群體的準確度偏差，較誼僅提供人口統計描述的作法更能控制代表性傷害，於「跨群體效能落差」與「偏差緩解紀錄」表現良好，惟其語料仍以英語及美國情境為主，「低資源語言覆蓋率」同樣偏低。此外，Ashokkumar 等人（2026）雖回報高度預測相關，但亦坦承模型有系統性高估效果量的傾向，恰可由「錯誤類型分析」指標加以標定。

上述回溯顯示，本框架具備實際的診斷能力：它能將兩個個案的方法優勢與盲點，轉化為可比較、可稽核的具體評分向度，使讀者得以即時檢驗既有研究的效度邊界，而不必等待完整的經驗驗證。此一演示亦印證，框架的價值不在於提供單一

分數，而在於促使研究者於研究設計階段即系統性地面對透明度、公平性與可重製性的取捨。

必須說明的是，本框架現階段仍屬規範性的建構，其完整效度仍有待後續經驗檢證。本文提出三階段驗證路徑：第一，以德菲法（Delphi method）凝聚社會科學家、資料科學家、研究倫理委員、期刊編輯對指標內容與權重的共識；第二，於政治學、社會學、傳播學與教育研究等子領域進行試點應用，檢驗各指標的可操作性與辨識度；第三，動態評估採用該框架前後，研究報告在透明度、可重製性與偏差揭露品質的變化。如此，本框架方能由本文所提出的輪廓，逐步發展為具實證支撐的治理工具。

為提升三階段驗證路徑的方法可複製性，本文進一步補充其操作程序。第一階段採德菲法，擬邀集 15 至 20 位異質專家，涵蓋社會科學方法學者、資料科學家、研究倫理委員會委員與期刊編輯，並力求各類人數均衡，以避免單一專業視角主導。程序包含二至三輪匿名問卷：第一輪以開放題蒐集指標項目與權重建議；第二、三輪則以李克特七點量表評定各指標的重要性與可操作性，並回饋群體統計結果供專家修正判斷。共識判準設定為單一指標四分位距（IQR）不大於一，或同意比例達 75% 以上；指標權重的一致性則以肯德爾和諧係數（Kendall's W）檢定，W 值達 0.7 以上即視為形成穩定共識（Hsu & Sandford, 2007）。

第二階段為子領域試點，擬於政治學、社會學、傳播學與教育研究各選取數篇近年應用 LLMs 的實證論文，由二至三名獨立評分者依框架指標評估，並以評分者間信度（如 Krippendorff's α 或組內相關係數）檢驗指標的可操作性與辨識度；α 值達 0.67 以上者視為可接受，未達標者則回饋修訂其操作型定義。同時，研究將記錄評分所需時間與爭議項目，以評估框架在實際審查情境中的可行性與執行成本。

第三階段採動態評估，擬以準實驗設計比較導入本框架前後的差異。具體而言，本文建議選定願意試行框架的期刊或研究團隊，蒐集其導入前後一定期間內研究報告在「資料來源揭露完整度」、「提示與版本登錄率」、「可重製性查核表完成度」與「偏差揭露品質」等指標上的變化，並以間斷時間序列或前後對照方式估計成效，作為後續修正與制度化的實證依據。

陸、結 論

本文以批判性文獻檢視與質性個案分析，檢視 AI 對社會科學研究、方法效度與倫理治理的影響。研究發現：AI 的重要性不僅有助於提升研究效率，更在於它將資料處理、概念測量、前導實驗、研究報告與責任歸屬重新連結，迫使社會科學重新檢視方法規範。

AI 的計算能力與分析效能正推動社會科學經歷根本性的範式改變。資料處理的自動化、模擬能力的革新、研究流程的優化與跨領域包容性的增進，這些機遇已從理論構想轉化為具體實踐，預示著未來社會科學研究的深刻變革。

不過，這些機遇的實現絕非自動導航。Argyle 等人（2023）的研究及史丹佛大學模擬人類受試者項目均顯示，AI 應用需要嚴謹的理論基礎、精細的方法設計與持續的倫理反思，更重要的是，機遇與挑戰是一體兩面的，提升效率的同時可能強化偏差，擴大研究規模可能降低深度反思。

在機會方面，AI 顯著提升資料分析的規模與速度，使研究者得以處理傳統方法無法觸及的複雜現象。生成式 AI 產生的合成資料集為理論測試與預測研究開闢新徑，自動化工具減少研究者的時間與心力，使其得以專注於更高階的理論建構與批判性詮釋。

在挑戰方面，偏差與公平性問題最為根本。訓練資料中的系統性偏差在模型輸出中獲得「科學化」的合法性，這對社會科學尤為危險，因為我們的研究結論直接影響公共政策與社會認知。模擬人類行為的侷限性揭示 AI 的本質限制，它無法真正複製人類行為的複雜性與多樣性，倫理、隱私與問責的缺失則反映 AI 治理框架的脫節，方法論的透明度與可重製性問題，更威脅著科學知識累積的基礎。

綜上，本文研究發現可歸納為：AI 對社會科學的影響並非機會與挑戰的簡單相加，而是雙重辯證的結構性結果，效率與規模的擴張伴隨效度與詮釋深度的耗損，自動化與生產力的提升，也可能弱化公平性與問責。

換言之，AI 確實重構資料標註、假設生成、前導模擬、分析詮釋與研究報告等環節，但其效率增益必須以效度驗證為前提。再者，模型偏差、提示敏感度、非確定性輸出與黑箱性質，威脅內在效度、外在效度與規範正當性。其次，透明度、公平性、可重製性與問責的操作化，提供研究過程中管理 AI 風險的可行路徑。

根據前述，本文提出以下建議：

(一) 建立負責任 AI 框架：社會科學研究機構應參考 OECD《人工智慧原則》、UNESCO《人工智慧倫理建議書》、Microsoft《負責任 AI 標準》、行政院（2023）《生成式 AI 指引》，以及教育部臺灣學術倫理教育資源中心（2024）強調的揭露、誠信、保密與責任原則，制定領域專屬規範。框架應涵蓋資料蒐集、模型選擇、提示紀錄、偏差審查、隱私保護、結果驗證與公開揭露，而非僅作為研究完成後的倫理聲明。

(二) 強化混合方法設計：鼓勵「人類研究者＋LLMs」分工，而非以 LLMs 取代人類樣本。具體而言，可先以 LLMs 進行前導模擬、題項測試與假設篩選，再以人類樣本、歷史資料或田野資料驗證結果。此設計可降低先導研究的投入成本，並防止將單一模型的輸出偏差誤認為可靠的實證結論。

(三) 提高透明度與可重製性：研究者應完整記錄模型名稱與版本、使用日期、確切提示、系統設定、抽樣參數、資料預處理流程、人工驗證比例與失敗案例。期刊與學會可制定 LLMs 應用報告準則，要求作者提交提示附錄、模型卡、資料表與可重製性查核表，以提升同儕審查品質。

(四) 投資偏差減輕與包容性工具：研究社群應優先建立涵蓋低資源語言、非西方文化脈絡與邊緣群體資料的驗證集，並發展分組評估、對抗性測試與在地化錯誤分析。包容性研究的重點不在讓模型替代弱勢群體發聲，而是運用模型辨識資料缺口，設計更具代表性的經驗研究。尤須強調臺灣在地脈絡的特殊性。由於繁體（正體）中文語料在全球訓練資料占比偏低，主流模型多以簡體中文語料為基礎，易將簡體語境的用語習慣、政治立場與歷史詮釋「二度抹平」臺灣的語言與社會結構（Lin & Chen, 2023）；同時，臺語（閩南語）、客語與原住民族語等本土低資源語言，普遍存在覆蓋不足與品質落差。因此，臺灣研究者導入 LLMs 時，宜將表 2 的「分組評估」指標，由既有的性別、族裔軸擴充納入「正體/簡體用語軸」與「族群語言軸」，並將「邊緣群體對照」的驗證集，明確納入原住民族、客家與新住民語料，以檢視主流模型是否系統性地邊緣化在地群體。此類在地化調整，應與行政院（2023）《生成式 AI 參考指引》、教育部（2024）之校園指引，以及國家層級的可信任 AI 對話引擎（如 TAIDE）等本土治理與技術資源形成對話，使負責任 AI 框架不僅移植西方原則，更能回應臺灣多語社會的具體處境。

(五) 促進跨學科與政策對話：社會科學家應積極參與 AI 治理與模型評估，與電腦科學、倫理學、法律、公共政策及受影響社群合作，將概念測量、制度脈絡與社會不平等的分析能力帶入 AI 發展與監管。

本文的貢獻可分為三個層面。理論上，本文以雙重辯證整合 AI 的方法增益與規範風險，避免將機會與挑戰視為彼此分離的面向。方法上，本文明確指出 AI 介入研究設計的關鍵節點，並區分前導模擬、資料標註、分析詮釋與結果報告各自的效度要求。實務與政策上，本文將負責任 AI 原則轉化為可檢核的資料表、模型卡、分組評估、提示登錄、可重製性查核表與隱私衝擊評估，作為社會科學研究機構與期刊制定規範的參考。

未來研究應進行經驗檢證，評估本文框架在政治學、社會學、傳播學、教育研究與公共政策評估等子領域的適用性，後續研究亦可比較不同模型、語言與資料類型的效度差異，並追蹤研究者使用 AI 後，在研究設計、寫作風格、引用實踐與學術誠信的長期變化。

總而言之，AI 不應被視為取代人類研究者的捷徑，而應被理解為需要制度化約束的研究基礎設施。社會科學若要在 AI 時代維持學術價值，重點不是拒絕技術，而是透過更嚴格的效度檢驗、透明報告與倫理問責，確保技術能力服務於公共知識與社會福祉，而非加劇既有不平等。

就研究方法而言，本文有三個限制：第一，本文未進行完整系統性回顧，文獻範圍可能受檢索詞與可得資料限制；第二，個案皆以英語語境與美國樣本為基礎，其結論未必直接適用於我國或其他非西方社會；第三，本文提出的框架仍待跨領域專家評估與實證檢驗，惟這些限制並不降低本文的研究價值，而是指出負責任 AI 研究治理需要持續累積經驗證據。就在地化而言，本文已於建議部分初步將西方治理原則與行政院（2023）、教育部（2024）之本土規範相互對照，並指出繁體中文與本土低資源語言的特殊風險，惟臺灣情境的完整經驗檢證，仍有待後續研究補足。

參考文獻

行政院（2023）。行政院及所屬機關（構）使用生成式 AI 參考指引。國家科學及技術委員會。https://www.nstc.gov.tw/folksonomy/list/c79bf57b-dc94-4aff-8d14-3262b5559cfc?l=ch

教育部臺灣學術倫理教育資源中心（2024）。大學校園因應生成式 AI 之指引及教學建議。https://ethics.moe.edu.tw/resource/epaper/html/21/

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202, pp. 337–371).

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669–674. https://doi.org/10.1126/science.adi6000

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2026). Large language models can predict the results of social science experiments. *Nature*, *656*(8126), 115–122. https://doi.org/10.1038/s41586-026-10742-x

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), Article e2314021121. https://doi.org/10.1073/pnas.2314021121

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics (pp. 5454–5476). https://doi.org/10.18653/v1/2020.acl-main.485

Chakravorti, T., Wang, X., Venkit, P. N., Koneru, S., Munger, K., & Rajtmajer, S. (2025). Social scientists on the role of AI in research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Chen, M. C. (2020). The impact of artificial intelligence on human society and bioethics. *Taiwanese Journal of Medical Ethics*, *32*(4), 339–343.

Davidson, T., & Karell, D. (2025). Integrating generative artificial intelligence into social science research: Measurement, prompting, and simulation. *Sociological Methods & Research*, *54*(3), 775–793.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Gefen, A. (2020). *AI for digital humanities and computational social sciences*. HAL Open Science. https://hal.science/hal-03043393/

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. 31122). https://www.nber.org/papers/w31122

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research, and Evaluation*, *12*(10), 1–8.

Ivanov, S. (2025). *Responsible use of AI in social science research*. ResearchGate. https://www.researchgate.net/publication/393910280_Responsible_use_of_AI_in_social_science_research

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kalhor, G., Ali, S., & Mashhadi, A. (2025). Measuring biases in AI-generated co-authorship networks. *EPJ Data Science*, *14*(1), 14–38.

Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T., & Yin, Y. (2025). Scientific production in the era of large language models. *Science*, *386*(6727), 1240–1245.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062.

Lin, Y.-T., & Chen, Y.-N. (2023). *Taiwan LLM: Bridging the linguistic divide with a culturally aligned language model* (arXiv:2311.17487). arXiv. https://arxiv.org/abs/2311.17487

Microsoft. (2022). *Microsoft responsible AI standard, v2.0*. https://msblogs.thesourcemediaassets.com/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Microsoft. (2023). *AI Forum 2023 | AI4Science: Accelerating scientific discovery with artificial intelligence*. https://www.microsoft.com/en-us/research/video/ai-forum-2023-ai4science-accelerating-scientific-discovery-with-artificial-intelligence

Microsoft. (2025). *Societal AI: Research challenges and opportunities*. https://www.microsoft.com/en-us/research/wp-content/uploads/2025/03/Societal-AI-Research-Challenges-and-Opportunities.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).

OECD. (2024). *OECD AI principles overview*. https://oecd.ai/en/ai-principles

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). *Generative agent simulations of 1,000 people* (arXiv:2411.10109). arXiv. https://arxiv.org/abs/2411.10109

Sarker, I. H., & Furht, B. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, *39*(1), 1–25.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2310.11324

Thapa, S., Shiwakoti, S., Shah, S., Adhikari, S., Veeramani, H., Nasim, M., & Naseem, U. (2025). Large language models (LLMs) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining*, *15*(1), Article 4. https://doi.org/10.1007/s13278-025-01428-9

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

UNESCO. (2024). *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.

Opportunities and Challenges of Artificial Intelligence for Social Science Research

William Yang*

Abstract

Artificial intelligence, particularly large language models (LLMs) and generative AI, has evolved from a research support tool into a methodological variable within the knowledge-production process of the social sciences. Adopting critical literature analysis as its primary approach, supplemented by qualitative case analysis, this article focuses on three research questions: first, which stages of social science research are substantively altered by LLMs and generative AI; second, how their reproduction of bias, simulation distortion, and black-box nature affect research validity and epistemic legitimacy; and third, what governance mechanisms can be operationalized throughout the research process. The findings indicate that the principal contributions of AI lie in enhancing the processing capacity for textual and unstructured data, supporting pilot virtual experiments and synthetic samples, and reducing the burden of data annotation and literature review. Its limitations, by contrast, primarily concern sample representativeness bias, high prompt sensitivity, the non-deterministic nature of outputs, as well as potential privacy risks and low reproducibility. This article proposes a Responsible AI application framework centered on transparency, fairness, reproducibility, and accountability, and translates it into auditable procedures including datasheets, model cards, disaggregated evaluation, prompt registration, and privacy impact assessment. The contribution of this article lies in consolidating the principled discussions dispersed across the literatures of AI ethics and computational social science into an operational analytical framework for social science research, thereby preventing researchers from privileging technical efficiency at the expense of methodological rigor.

Keywords: Artificial Intelligence, Social Science Research, Large Language Models, Computational Social Science, Responsible AI

* William Yang received his Ph.D. in Political Science and International Relations from University of Warwick (U.K.) and is currently Professor in the Department of Public Administration and Management at Shih Hsin University. He has previously served as Dean of International Affairs, Dean of Public Affairs, and Chair of the General Education Committee. His research expertise includes AI in Social Sciences, Trumpism, Global Governance, International Climate Politics. Email: william@mail.shu.edu.tw.
The paper was published under two double-blind reviews. Received: July 28, 2026. Accepted: August 16, 2026.