



贵州民族大学学报(哲学社会科学版)

Journal of Guizhou Minzu University(Philosophy and Social Sciences)

ISSN 1003-6644,CN 52-1155/C

## 《贵州民族大学学报(哲学社会科学版)》网络首发论文

题目：生成式人工智能的兴起对社会科学研究方法的拓展——以偏见研究为例  
作者：卢云峰，黄星尧，吴语菲，周旅军  
网络首发日期：2025-03-27  
引用格式：卢云峰，黄星尧，吴语菲，周旅军. 生成式人工智能的兴起对社会科学研究方法的拓展——以偏见研究为例[J/OL]. 贵州民族大学学报(哲学社会科学版), <https://link.cnki.net/urlid/52.1155.c.20250327.1042.002>



**网络首发：**在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

**出版确认：**纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

# 生成式人工智能的兴起对社会科学研究方法的拓展

——以偏见研究为例

卢云峰,黄星尧,吴语菲,周旅军

**摘要:**生成式人工智能(Generative Artificial Intelligence)技术的崛起,为社会科学研究方法的扩展带来了契机。基于变换器(Transformer)架构的大语言模型由于经过大量预训练能够更有效地理解和处理复杂的上下文信息,吸引了众多社会科学家的关注。社会科学家对生成式人工智能应用的探索,一方面集中于其在数据分析、充当实验中介和协助写作等研究过程中的效果;另一方面关注生成式人工智能所具有的模拟人类对象的能力。论文以偏见研究为具体案例,从“生成式人工智能作为研究工具”和“生成式人工智能作为研究对象”两个方面回顾和探讨了社会科学如何调整自身的研究实践以面对大语言模型应用的趋势,并总结生成式人工智能给社会科学研究方法带来的助益。

**关键词:**生成式人工智能;方法论;社会科学;DeepSeek;ChatGPT

**中图分类号:**C34 **文献标识码:**A **文章编号:**1003-6644(2025)03-0000-22

作者卢云峰,男,博士,贵州民族大学社会学院特聘教授(贵州 贵阳 550025),北京大学社会学系教授、博士生导师(北京 100871);黄星尧,男,北京大学社会学系博士生(北京 100871);吴语菲,女,北京大学社会学系博士生(北京 100871);周旅军,男,博士,中华女子学院副教授(北京 100101)。

社会科学研究方法始终与技术进步密切相关。工业革命之后,英国和德国率先开展了对工人和农民的大规模社会调查,收集有关社会变迁的基础数据;法国年鉴学派将计量方法应用于历史研究,极大影响了阿道夫·凯特莱(Adolphe Quetelet)从统计的

\* 本文为北京大学武汉人工智能研究院开放课题“数字社会的核心特征与治理模式研究:以武汉光谷为例”阶段性成果。

平均值和正态分布角度理解社会规律的思路。<sup>①</sup>随后,实验法与抽样调查相继推动了社会科学研究方法的发展与完善;计算机技术的飞跃则进一步为分析人类行为、态度和交互提供了更强大的算力和更庞大的数据集。<sup>②</sup>如今,生成式人工智能(Generative Artificial Intelligence,简称 GAI)技术的崛起在社会各界再次引发讨论,2025 年 1 月发布的中国大语言模型 Deepseek 更是凭借其卓越的性能和显著的成本优势,将这一波技术变革的势能推向了新的高峰。<sup>③</sup>在此背景下,讨论如何在社会科学研究中合理地利用生成式人工智能显得尤为必要。

实际上,社会科学家已经在此领域进行了不少探索,尤其关注基于变换器(Transformer)架构的大语言模型(Large Language Models,简称 LLMs)。这类模型经过海量数据预训练,能够更有效地理解和处理复杂的上下文信息。目前的研究主要沿着两个方向展开:一是生成式人工智能的工具性应用研究,包括数据分析、实验中介和写作辅助等。比如,赫赛尔廷(Michael Heseltine)与冯·霍恩堡(Bernhard Clemm von Hohenberg)证明了 GPT-4 对政治新闻在情感、意识形态等变量上进行的编码具有和人工专家类似的准确性。<sup>④</sup>二是模拟人类对象的研究,如孙胜钟(Sun Seungjong)等人发现大语言模型能够模拟以群体为单位的美国民意调查结果。<sup>⑤</sup>

① Lazarsfeld P. F. , "The Sociology of Empirical Social Research" , *American Sociological Review* , vol. 27 , no. 6 , 1962.

② Webster M. , & Sell J. , "Theory and Experimentation in the Social Sciences" , in Outhwaite , William and Turner , Stephen P. , ed. , *The SAGE Handbook of Social Science Methodology* , London : SAGE Publications , 2007 , pp. 190 - 207 ; Osborne T. , & Rose N. , "Do the Social Sciences Create Phenomena? : the Example of Public Opinion Research" , *The British Journal of Sociology* , vol. 50 , no. 3 , 1999 ; Meyer E. T. , & Schroeder R. , *Knowledge Machines : Digital Transformations of the Sciences and Humanities* , Cambridge , MA : The MIT Press , 2015.

③ "How China Created AI Model DeepSeek and Shocked the World" , 2025 - 1 - 30 , <https://www.nature.com/articles/d41586-025-00259-0> , 2025 - 2 - 10.

④ Heseltine M. , & Clemm von Hohenberg B. , "Large Language Models as a Substitute for Human Experts in Annotating Political Text" , *Research & Politics* , vol. 11 , no. 1 , 2024.

⑤ Sun S. , Lee E. , Nan D. , et al. , "Random Silicon Sampling: Simulating Human Sub - population Opinion Using a Large Language Model Based on Group - level Demographic Information" , 2024 - 2 - 28 , <https://doi.org/10.48550/arXiv.2402.18144> , 2025 - 1 - 28 ; Kim J. , et al. , "Learning to be Homo Economicus: Can an LLM Learn Preferences from Choice" , 2024 - 1 - 14 , <https://doi.org/10.48550/arXiv.2428> ; Suri , 01.07345 , 2025 - 1 - G. , et al. , "Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT - 3.5" , *Journal of Experimental Psychology: General* , 2024.

这些新的发展无疑向社会科学提出了拷问:我们应该如何更好地适应大语言模型与学科结合的趋势? 本文将以偏见研究为具体案例,从“生成式人工智能作为研究工具”和“生成式人工智能作为研究对象”两个方面来回顾和探讨生成式人工智能带给社会科学研究方法的助益和风险。传统的偏见研究主要探讨人们在认知、态度或行为上对某种群体的负面取向。<sup>①</sup>一方面,生成式人工智能可能作为研究工具参与到传统的人类偏见研究中的各个环节,显示出其对社会科学研究方法的潜在助益;另一方面,带有偏见的生成式人工智能本身可以成为研究对象,因此计算机领域研究 GAI 偏见的具体技术可以为社会科学探索新的研究方法带来启发。

## 一、作为研究工具的生成式人工智能

“作为研究工具的生成式人工智能”是指它作为客体帮助研究者认识世界,参与研究各个具体环节的功能。<sup>②</sup>朗克尔(Runkel P. J.)和麦格拉思(McGrath J. E.)曾根据研究操作的阻塞性(obtrusive/unobtrusive research operations)和行为系统的普遍性(universal/particular behavior systems)将传统研究方法分为四类:(1)实验法、判断任务(强阻塞性、普遍性);(2)抽样调查、形式理论(低阻塞性、普遍性);(3)实验刺激、实地实验(强阻塞性、特殊性);(4)田野调查、计算机模拟(低阻塞性、特殊性)。它们存在一种三角困境,即不能同时保证跨行为者发现的普遍性、测量的控制和精度,以及情境的现实性。<sup>③</sup>近几十年来,计算社会科学对大数据的利用力求应对这一困境,形成数据驱动而非完全理论驱动的“第四范式”。<sup>④</sup>生成式人工智能的加入既可能提升传统研究模式的效率,从而加速第四范式的进程。

偏见指人类对非自己所属群体持有的总体负面态度,与对某种群体形成的某个特

① Dovidio J., Miles Hewstone, Peter Glick, et al., "Prejudice, Stereotyping and Discrimination: the Theoretical and Empirical Overview", in Dovidio J., et al., ed., *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, London: SAGE Publications Ltd, 2010.

② Xu R., Sun Y., Ren M., et al., "AI for Social Science and Social Science of AI: A Survey", *Information Processing & Management*, vol. 61, no. 3, 2024.

③ Runkel P. J., & McGrath J. E., *Research on Human Behavior: A Systematic Guide to Method*, New York: Holt, Rinehart and Winston, 1972.

④ Lazer D., Pentland A., Adamic L., et al., "Computational Social Science", *Science*, vol. 323, no. 5915, 2009.

征的特定印象(即刻板印象, stereotype)以及基于此付诸的行为(即歧视, discrimination)密切相关。<sup>①</sup>本文将上述三类研究都视为偏见研究。长期以来,社会学、心理学、政治学等学科对其进行了大量的研究,内容涉及偏见的类型、偏见的形成过程、偏见的影响、偏见的干预等多个方面。<sup>②</sup>

偏见研究的传统数据收集方法以问卷调查、量表填写和实验法为主。新趋势下,众多计算社会科学家开始利用在线数据、实验、虚拟现实技术和代理建模的方式研究有关政治观点极化的问题,试图贴近深刻受到互联网影响的当代偏见现象。<sup>③</sup>本文将从传统研究范式涉及的步骤(理论到假设、数据收集和数据分析)以及第四范式涉及的操作过程来讨论生成式人工智能对社会科学研究方法的拓展以及在偏见研究中的应用潜力。

## (一) 传统研究范式

### 1. 从理论到假设

生成式人工智能是否能激发研究者的理论灵感?答案是可能的。首先,包括 GPT 系列和 Claude 在内的大语言模型能够帮助研究者进行针对某一特定问题的头脑风暴。计算社会学家克里斯托弗·拜尔(Christopher A. Bail)尝试让生成式人工智能针对“社交媒体中的政治两极分化”现象生成研究问题。尽管回答质量参差不齐,但仍涌现出一些有启发性的问题,比如“为什么人们会迁移社交媒体平台,以及这种迁移对政治两极分化的影响机制”<sup>④</sup>。其次,大语言模型对于特定论点的追问也有助于探索现有理论的缺陷。有研究团队以 DeepSeek V2 - Chat 为核心语言模型构建了一个有启发式提问

① Kite M. E. , Bernard E. Whitley, Jr. , Lisa S. Wagner, *Psychology of Prejudice and Discrimination* , New York: Routledge, 2022.

② Allport G. , *The Nature of Prejudice* , Reading, MA: Addison - Wesley, 1954; Blumer H. , " Race Prejudice as a Sense of Group Position" , *the Pacific Sociological Review* , vol. 1 , no. 1 , 1958; Huddy L. , & Feldman, S. , " On Assessing the Political Effects of Racial Prejudice" , *Annual Review of Political Science* , vol. 12 , no. 1 , 2009.

③ Edelman A. , Wolff T. , Montagne D. , et al. , " Computational Social Science and Sociology" , *Annual Review of Sociology* , vol. 46 , no. 1 , 2020.

④ Bail C. A. , " Can Generative AI Improve Social Science? " , *Proceedings of the National Academy of Sciences* , vol. 121 , no. 21 , 2024.

功能的研究助手<sup>①</sup>,通过引导用户逐步明确研究问题,显著提升了研究效率。

为了直观说明生成式人工智能在演绎推理阶段对社会科学研究的辅助作用,我们以社会认同理论为例,令生成式人工智能进行反思;在学习了有关社会认同理论的基本材料后,GPT-4 所给出的三个反驳论点中有两个十分有力。例如,“数字时代,人们可以轻易地改变或隐藏自己的社会身份,参与在线社群,这些社群往往超越了地理和社会界限。这种匿名性和流动性挑战了理论中关于固定社会分类和群体忠诚的假设”<sup>②</sup>。不过,生成式人工智能作为研究助手的成熟度仍需进一步探讨。它能在多大程度上协助理论创新,实际上取决于它是否具备反事实思考、创造力、溯因推理等理论化(theorizing)能力。<sup>③</sup>随着智能技术的发展,社会科学领域常见的跨学科概念启发、寻找理论无法解释的新现象,以及在全新的概念空间当中理解世界等实现理论创新的方法,与计算机实际上有能力达到的组合型创造力(combinationl creativity)、探索性创造力(exploratory creativity),以及转化型创造力(transformationl creativity)存在建立连接的可能。<sup>④</sup>

## 2. 数据收集

生成式人工智能对数据收集的优化作用体现在四个方面。首先,在经典实验研究范式中,生成式人工智能有可能改善困扰实验法的外部有效性问题。具体而言,该技术可以有效应对现实性与可控性挑战:前者涉及传统实验中被试群体的同质化特征,后者源于实验效应等因素的干扰。研究表明,大语言模型在经过算法调整和语言阐释监督后能够提升自身的文化适应性<sup>⑤</sup>,已经有众多研究者将经典实验研究迁移到了大

① Zheng Y., Sun S., Qiu L., et al., "Open Researcher: Unleashing AI for Accelerated Scientific Research", 2024-8-13, <https://doi.org/10.48550/arXiv.2408.06941>, 2025-1-7.

② 我们所给予的提示(prompt)如下:你是一名专业的社会心理学研究者,请阅读给你的文件,它是一篇有关社会认同理论的简介(参见 Hornsey M. J., "Social Identity Theory and Self-categorization Theory: A Historical Review", *Social and Personality Psychology Compass*, vol. 2, no. 1, 2008)。请结合当代现实,给出反驳社会认同理论三个论点。

③ Felin T., & Holweg M., "Theory Is All You Need: AI, Human Cognition, and Causal Reasoning", *Strategy Science*, vol. 9, no. 4, 2024.

④ Boden M., *The Creative Mind*, New York: Routledge, 2004.

⑤ McIntosh T. R., Liu, T., Susnjak, T., et al., "A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination", *IEEE Transactions on Artificial Intelligence*, 2023.

模型之上(例如,内隐测量、最后通牒博弈)<sup>①</sup>。

其次,在涉及复杂伦理问题的场合,生成式人工智能使社会科学家有可能在减少风险的情况下探索社会偏见这样的敏感议题。犹太裔心理学家罗伯特·扎荣茨(Robert B. Zajonc)提出,群际偏见是一个包括“极端化的社会分类”“道德排斥”等复杂环节的社会过程。<sup>②</sup>他所依据的材料与种族和宗教大屠杀有关,现在已经不可能对其时代背景进行复现。如果要在一个考虑到文化敏感性和历史复杂性的框架内研究扎荣茨提到的现象,那么生成式人工智能最有可能提供跨文化的实验设计以及自然的实验情景模拟。通过模拟实验和创建虚拟道德决策情景,避免将参与者置于可能造成心理或情感伤害的环境中。<sup>③</sup>

再次,生成式人工智能可以与智能体仿真协同发挥效力,探索复杂现象涌现机制。智能体仿真,例如基于主体建模(agent-based modeling)的方法,近年来在社会科学领域获得广泛运用。仿真模拟法将社会视为一个复杂系统,其核心特征包括非线性反应、反馈循环、自组织行为,以及出现从微观行为到宏观模式的涌现现象。<sup>④</sup>它有三个核心构念:(1)主体,指能够感知环境、做出决策、采取有意图行动的智能体行动者;(2)环境,指主体所处的场域;(3)规则,即嵌入主体的行为导向与限制。<sup>⑤</sup>

现有研究多将生成式人工智能的自主学习、情境分析、自然语言处理能力整合到仿真模拟的主体中,为社会偏见的机制研究提供新的可能。在主体异质性方面,生成式人工智能能够精准捕捉现实社会个体在社会、经济、心理和人口等多维背景下的差

① Steed R. & Caliskan A. , "Image Representations Learned with Unsupervised Pretraining Contain Human-like Biases" , *Conference on Fairness, Accountability, and Transparency*, 2021; Abramski K. , Citraro S. , Lombardi L. , et al. , "Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students" , *Big Data and Cognitive Computing*, vol. 7, no. 3, 2023; Guo F. , "GPT in Game Theory Experiments" , 2023-5-9, <https://doi.org/10.48550/arXiv.2305.05516>, 2025-1-7.

② Zajonc, R. B. , "The Zoomorphism of Human Collective Violence" , In Newman L. S. , Erber R. , *Understanding Genocide: The Social Psychology of the Holocaust*, Oxford: Oxford University Press, 2002.

③ Dillion D. , Tandon N. , Gu Y. , "Can AI Language Models Replace Human Participants?" , *Trends in Cognitive Sciences*, vol. 27, no. 7, 2023.

④ 吕鹏、陈典涵:《社会复杂系统智能模拟:涌现机理与方法路径》,《山东大学学报(哲学社会科学版)》2024年第1期。

⑤ 吕鹏、范晓光:《计算社会科学导论》,北京:清华大学出版社,2023年。

异,批量生成异质性主体,从而提升预测和分析的准确性。<sup>①</sup>另外,在主体对环境的适应上,大模型对外部环境的开放性允许其认识到自身的可供性,从而主动感知调度,在零成本训练的前提下自适应各种环境。<sup>②</sup>最后,在主体间互动上,得益于较好的语义理解能力(如理解复杂的自然语言结构、语境、双关语,以及隐含的意义等),智能体间不仅能进行自然语言交互,还能深入理解和解释环境中的信息,实现“智能沟通”(intelligent communication)。<sup>③</sup>总而言之,生成式人工智能对计算社会学经典范式的拓展,体现为建构了一个兼具高可控性、微观可解释性、时变特征与社会复杂性的数字孪生系统。

最后,在大规模社会调查中,生成式人工智能有望在问卷编制上减轻人为偏见。由于问卷和量表过往依赖于人工编制,在产生这类研究工具的过程中很有可能引入人为偏见,带来了诸如早期种族主义量表当中的适应性问题。<sup>④</sup>生成式人工智能协助参与编制的量表则有可能改善这一问题。已有研究证明由 GPT-4 参与开发的量表具有较高的信度和效度,并认为在结合专家知识的情况下相比于传统量表开发在改善人为偏见上具有很大的潜力。<sup>⑤</sup>

### 3. 数据分析

生成式人工智能在数据分析方面的优势目前主要集中于非结构化数据编码,包括处理数据的量级以及更加精确的语义理解这两个方面。传统的定性数据编码通常依靠众包人工标注或早期深度学习模型,但前者成本较高、不同标注者之间标准化程度低,后者处理长文本时准确率低、缺乏语境理解、需要大量针对性训练。生成式人工智能出色的语义理解能力使其能够识别细腻和具有情境依赖性的情感。变

① Hommes C. & LeBaron B. , *Computational Economics: Heterogeneous Agent Modeling*, Amsterdam: North Holland, 2018.

② Xi Z. , Chen W. , Guo X. , et al. , "The Rise and Potential of Large Language Model Based Agents: A Survey" , *Science China Information Sciences*, vol. 68, no. 2, 2025.

③ Gu Z. , Zhu X. , Guo H. , et al. , "AgentGroupChat: An Interactive Group Chat Simulacra For Better Eliciting Collective Emergent Behavior" , 2024 - 3 - 20, <https://doi.org/10.48550/arXiv.2403.13433>, 2025 - 1 - 10.

④ Fiske S. T. & North M. S. , "Measures of Stereotyping and Prejudice: Barometers of Bias" , in Boyle G. J. , Saklofske, D. H. , & Matthews, G. (Eds. ) , *Measures of Personality and Social Psychological Constructs*, Academic Press, 2014.

⑤ Mohammed Salah A. , Abdelfattah, F. , Al Halbusi H. , et al. , "Can Generative AI Craft Scale Items? A Mixed - Method Study on AI's Capability to Adapt and Create New Scales with Recommendations for Best Practices" , *SSRN Preprint*, 4931424, 2023.

换器架构中的自注意力(self-attention)机制帮助模型有效处理长距离的数据依赖;多头注意力(multi-head attention)允许模型从多角度学习信息;词嵌入(word embedding)则将文字转换成机器能理解的向量形式。<sup>①</sup>这些技术共同提升了大模型处理非结构化材料的能力。加米尔迪安(Yasir Gamielien)等人将GPT-3.5主题分析的结果与3,800名学生手工编码的结果进行了比对,发现GPT-3.5具有优秀的生成高颗粒度编码的能力。<sup>②</sup>科兹洛夫斯基(Austin C. Kozlowski)等人使用传统词嵌入模型Word2Vec来对“社会阶层”这一概念在20世纪的变迁进行深入分析,比较发现机器编码有效地克服了传统文本分析中阐释的主观性以及无法对词间关系进行更细致处理的问题。<sup>③</sup>

## (二) 第四范式

受到托马斯·库恩(Thomas Kuhn)的范式理论启发,吉姆·格雷(Jim Gray)提出了人类认知的四次范式革命:伽利略时代之前的实验科学为第一范式、牛顿之后的理论科学为第二范式、计算仿真技术支持的计算科学为第三范式,以及以数据密集型发现为主导的第四范式。<sup>④</sup>社会科学与自然科学研究方法的演进是同步的——在社会科学中,这四种范式分别对应经验观察(以田野调查为典型)、理论建构、计算社会科学,以及当下深度融合人工智能的数字化社会科学。

阿盖尔(Lisa P. Argyle)等人在2023年的著名研究中提出“硅采样”(silicon sampling)社会调查,将第四范式方法论具象化。其核心机制在于:对GPT-3大语言模型进行提示工程,构建具有社会人口统计学特征的条件化生成模型,进而模拟特定

① Vaswani A., Shazeer N., Parmar N., et al., "Attention is all you need", *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, 2017.

② Gamielien Y., Case, J. M., & Katz, A., "Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding", *SSRN preprint*, 4487768, 2023.

③ Kozlowski A. C., Taddy M., & Evans J. A., "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings", 2018 - 3 - 25, <https://doi.org/10.1177/0003122419877135>, 2025 - 1 - 10.

④ Hey T., Tansley S., & Tolle K., "Jim Gray on eScience: A Transformed Scientific Method", in Hey T., Tansley S., & Tolle K., *The Fourth Paradigm*, Redmond, Washington: Microsoft Research, 2009, pp. xxvii - xxxi.

人类亚群体的意见、偏见和投票模式。通过算法保真度 (algorithmic fidelity) 验证, 硅基样本的输出不仅与人类样本的响应分布具有统计一致性, 更因其非人特性有效规避了社会期望偏差。<sup>①</sup> 该研究体现了第四范式的三重突破性特征。第一, 基于海量数据。依托 1,750 亿参数与 45TB 训练数据, GPT-3 突破了传统研究工具的信息处理体量边界。第二, 研究步骤不再受制于科学轮的循环过程。开展问卷调查之前, 研究者既可以延续传统, 用理论推演出研究假设, 还能让硅基样本大规模预填写问卷, 并从数据中发掘出规律, 帮助形成研究假设。由此在假设提出阶段也可以融合归纳逻辑与演绎逻辑的双向路径。第三, 推动跨学科协同创新模式。生成式人工智能的低成本、高可复现性, 意味着基于该技术的研究过程与成果可以被跨学科共享。为公共舆论研究、政治行为研究等学科互动性较强的社会科学领域提供了重要智识资源。

## 二、作为研究对象的生成式人工智能

“作为研究对象的生成式人工智能”指围绕生成式人工智能所进行的一系列有关机器行为或心理学 (machine psychology) 的研究, 它包括研究大语言模型所具有的偏见、歧视, 以及输出有毒内容等现象。<sup>②</sup> 过往的偏见研究通常基于人类主体的预设, 但与生成式人工智能有关的大量研究证实, 机器同样可以成为道德主体和道德代理人。<sup>③</sup> 值得注意的是, 社会科学研究将生成式人工智能作为研究对象时, 本质上仍以理解人类社会为终极目标。在这种意义上, 生成式人工智能既是待解析的技术实体, 又是折射现实社会的棱镜工具。

在此, 本文重点关注基于提示符的探测法 (prompt-based probing), 这种方法通过精心设计的提示来揭示模型在处理不同类型的输入时可能存在的不公正或偏见, 已经

① Argyle L. P., Busby E., Fulda N., et al. "Out of One, Many: Using Language Models to Simulate Human Samples", *Political Analysis*, vol. 31, no. 3, 2023.

② Mitchell M., & Krakauer D. C., "The Debate Over Understanding in AI's Large Language Models", *Proceedings of the National Academy of Sciences*, vol. 120, no. 13, 2023.

③ Bonnefon J. F., Rahwan I., & Shariff, A., "The Moral Psychology of Artificial Intelligence", *Annual Review of Psychology*, vol. 75, no. 1, 2024.

被广泛用于对生成式人工智能内生偏见的研究。<sup>①</sup>值得注意的是,使用这种方法时,评估数据和训练数据的相关性会导致评估时出现领域重叠、答案泄露和知识覆盖等问题,最终有可能影响评估的公正性和准确性。<sup>②</sup>因此,设计一个关于探测法的原创性基本框架并使之区别于目标大模型的预训练数据库,是有必要的。我们结合罗伯特·斯滕伯格(Robert J. Sternberg)的成功智力理论(*theory of successful intelligence*)<sup>③</sup>和弗朗索瓦·乔莱特(François Chollet)的人工智能研究设计了一个以提示符探测生成式人工智能偏见的基本框架。

成功智力理论将人类智力拆解为三个方面:(1)分析性智力(analytical intelligence),传统智商测试所衡量的智力,涉及问题解决、判断和评估等能力。分析性智力使人能够完成学术、问题解决和其他类似的智力活动。(2)创造性智力(creative intelligence),这种智力涉及创造、发现新问题和看待事物的新方法。创造性智力使人能够在面对新奇情况时适应和创新。(3)实用性智力(practical intelligence),又称为“街头智慧”,这种智力涉及日常环境中的任务,如在特定情境中有效地使用信息。实用性智力有助于个体适应、塑造或选择环境。<sup>④</sup>

在此基础上,乔莱特则指出人工智能的智力关键在于“技能获取效率”,即系统在不同任务和环境中学习和适应的能力。他给出了一个旨在测试系统面对未知和抽象任务时表现的框架,强调一个高智能的系统应该能够在较少经验和先验知识的情况下,解决具有较高泛化难度的问题。<sup>⑤</sup>受到乔莱特的启发,本文将斯滕伯格的成功智力

① Abid A. ,Farooqi M. ,& Zou J. , "Persistent Anti - Muslim Bias in Large Language Models" , *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* ,2021 ;Choudhary T. , "Reducing Racial and Ethnic Bias in AI Models: A Comparative Analysis of ChatGPT and Google Bard" ,*Proceedings of the 36th International RAIS Conference on Social Sciences and Humanities* , 2024.

② Cao Y. ,Li S. ,Yan Z. ,et al. , "A Comprehensive Survey of AI - Generated Content ( AIGC ): A History of Generative AI from GAN to ChatGPT" ,2023 - 3 - 7 ,<https://doi.org/10.48550/arXiv.2303.04226> ,2025 - 1 - 11.

③ Sternberg R. J. , *Wisdom, Intelligence, and Creativity Synthesized* , Cambridge : Cambridge Press , 2003 ,pp. 46 - 62

④ Sternberg R. J. , "The Theory of Successful Intelligence" ,*Review of General Psychology* ,vol. 39 , no. 2 ,2005.

⑤ Chollet F. , "On the Measure of Intelligence" ,2019 - 11 - 25 ,<https://doi.org/10.48550/arXiv.1911.01547> ,2025 - 1 - 11.

论迁移到生成式人工智能的框架中,从分析性智力、创造性智力和实用性智力三个方面来设立有关探测生成式人工智能偏见的基本框架(见图1)。

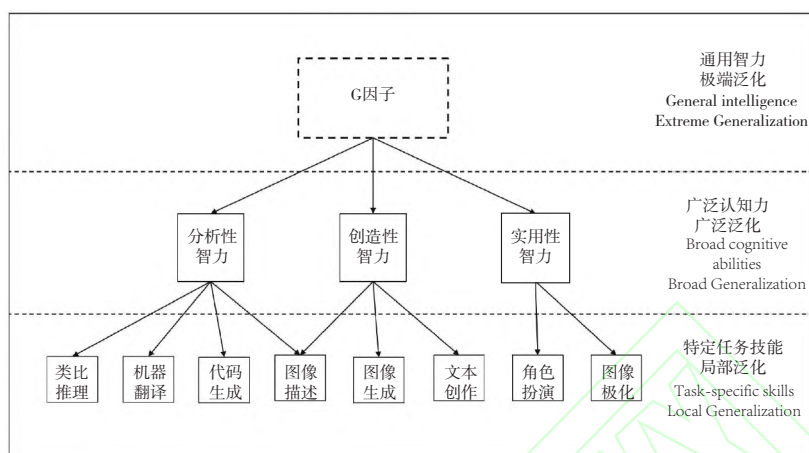


图1 探测生成式人工智能偏见的任务框架<sup>①</sup>

该理论框架的解释力基于一个前提:人类智力理论在人工智能领域也具有适配性。创造性、分析性、实用性智力在本质上是对认知过程的功能性划分,这与生成式AI的文本生成、逻辑推理、情境适应三大核心能力形成结构对应。分析性智力对应的类比推理任务则可检测逻辑规则中的归因偏见;创造性智力对应的图像/文本生成任务可揭示数据中潜藏的文化原型偏差;实用性智力对应的情境化创作任务能够评估系统对复杂社会语境的适应性,反映技术系统在价值排序中对特定文化的隐性倾斜。

图1中最底层为代表特殊智力技能的基础任务,包括但不限于类比推理、机器翻译、代码生成、图像描述、图像生成、文本创作、角色扮演,以及图像极化。其中有的任务可能对应多项智力类型。实际上,这些基础任务可以自由组合成综合任务,以实现刺激-推断和推断-刺激的双向任务考察。<sup>②</sup>下文将简要介绍每个基础任务的过程和特点,并对其偏见探测效果进行评价。

① 改编自乔莱特。在该框架中,人类智力结构有三层:顶部的通用智力(g因子),中间的广泛认知能力,以及底部的专业技能或测试任务。本文参考这一框架将人工智能智力也分为三层:“局部泛化”(从系统处理来自已知分布的新数据点的能力)、“广泛泛化”(系统在没有进一步人为干预的情况下处理广泛类别任务和环境的能力),与作为目标的、仅有生物形式的智能具有的“极端泛化”(具有处理完全新任务的开放式系统的能力)。参见 Chollet F., "On the Measure of Intelligence", 2019-11-25, <https://doi.org/10.48550/arXiv.1911.01547>, 2025-1-11.

② Kamruzzaman M., Shovon, M. M. I., & Kim G. L., "Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models", 2023-9-16, <https://doi.org/10.48550/arXiv.2309.08902>, 2025-1-11.

## (一) 分析性智力任务

分析性智力通常关注智能体判断、辨别、寻找问题答案等逻辑相关的能力。在生成式人工智能的偏见检测框架中,分析性智力维度下的测试之所以能够有效捕捉逻辑规则中的归因偏见,源于其在认知映射过程中对因果结构与属性关联模式的强制显性化。它所牵涉的具体生成任务包括类比推理、机器翻译、代码生成和图像描述等。

### 1. 类比推理

类比推理(analogical reasoning)涉及在两个或多个不同对象、事件或概念之间找到关系相似性。<sup>①</sup>侯世达(Douglas Richard Hofstadter)和伊曼纽尔·桑德尔(Emmanuel Sander)在著作《表象与本质:类比,思考之源和思维之火》中论及,类比是认知的基本单位。人类通过不断将新事物与既有经验类比而建立关联,是为认识世界的途径。每个概念都可以被看作一簇类比网络,对判断和决策具有引导作用。所以,类比推理的结构能够探测人类或人工智能思维中存在的认知图式和逻辑架构。人工智能领域自20世纪60年代开始对类比推理过程的机器模拟进行研究,这个过程通常涉及两对元素:源(source)和目标(target)。源是已知的关系或概念,而目标是需要应用这种关系或概念的新场景。在生成式人工智能的偏见识别任务中,类比推理一般表现为给出一组包含关键概念和空白位置的陈述或问题,并让大语言模型给出其认为与关键概念类似的词语填补空白位置。

波鲁巴斯(Tolga Bolukbasi)等人使用类比推理的方法,系统测量了词嵌入中存在的性别偏见。当输入特定性别关联种子词对(如“男性-女性”)时,大模型会输出具有性别刻板印象的职业配对[如“程序员-家庭主妇(homemaker)”。<sup>②</sup>这意味着性别中立的职业术语在向量空间中映射出系统性的语义偏移。因为词嵌入的语言结构是通过学习语料库而获得,所以大模型输出的社会偏见并非幻觉,而很大程度上是被统计学习模型吸收和放大的真实偏见。

<sup>①</sup> Gentner D. , " Structure - mapping: A Theoretical Framework for Analogy " , *Cognitive Science* , vol. 7 , No. 2 , 1983 .

<sup>②</sup> Bolukbasi T. , Chang K. , Zou J. , et al. , " Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings " , *Advances in Neural Information Processing Systems* , 29 , 2016 .

## 2. 机器翻译

机器翻译技能早在在大语言模型出现以前就已得到充分开发,该任务指将文本从一种自然语言翻译成另一种自然语言。

戈什(Souroj Ghosh)和卡利斯克坎(Aylin Caliskan)研究了 ChatGPT 在进行英语与孟加拉语(以及其他五种使用性别中立代词的低资源语言)之间的翻译时如何处理性别代词。他们发现,ChatGPT 常受到职业性别刻板印象的影响,将性别中立的原代词翻译为“他”或“她”。<sup>①</sup>

偏见问题(如性别偏见)具有文化差异,而过往的众多研究均将其视为同质的问题加以研究。<sup>②</sup>生成式人工智能带来了一个机会,让社会科学研究者能够以更加具有文化多样性的人工智能协助探索机器翻译中的偏见问题。

## 3. 图像描述

图像描述任务中,用户向智能模型提供目标图像;计算机通过对色彩、形状、符号的视觉识别,辅以自然语言处理系统,生成人类可读、可评价的描述性文本。图像描述是图片识别的进阶任务,不仅需要识别图中物体,还对图文信息的整合、抽象符号与现实意义的融贯有更高的要求。

萨尔罕(Habiba Sarhan)与海格里希(Simon Hegelich)从新闻网站上搜集了1000张具有典型政治意义的图像,让微软云计算服务(2021年12月公开版本)提供图片说明文本。研究发现,智能模型因为刻意回避带刻板印象的描述,反倒丧失了政治敏锐性,无视了弱势群体的不利处境,延续了话语秩序中的结构性压迫。例如,“在街道上席地休息的劳累移民”被人工智能描述为“躺在沙滩上”,其采取的“无偏中立”语言实则默认了世界北方(Global North)的奢侈生活方式。<sup>③</sup>

① Ghosh S. , & Caliskan A. , " ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non - gendered Pronouns : Findings Across Bengali and Five Other Low - resource Languages " , *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* , 2023 , pp. 901 - 912 .

② Pikuliak M. , Hrkova A. , Oresko S. , & Šimko M. , " Women are Beautiful , Men are Leaders : Gender Stereotypes in Machine Translation and Language Modeling " , 2023 - 12 - 30 , <https://doi.org/10.48550/arXiv.2311.18711> , 2025 - 1 - 11 .

③ Sarhan H. , & Hegelich S. , " Understanding and Evaluating Harms of AI - generated Image Captions in Political Images " , *Frontiers in Political Science* , vol. 5 , 2023 .

由于大语言模型主要通过预测下一个词元(token)来学习语言知识,而这种训练目标并不直接对应图像理解所需的特征提取和模式识别能力。人工智能在处理图像时会将其转换为序列形式输入,加之图像—文本的训练数据通常比纯文本数据的质量更难控制,因此,这种转换过程增加了模型暴露潜在偏见的机会。

#### 4. 代码生成

代码生成的任务是一种把程序员自身思路实现为应用软件的过程,需要有意调用数学思维、语言能力等高级认知功能(higher cognitive functions)与问题解决、逻辑规划等高级执行管理功能(higher-level executive functions)。<sup>①</sup>但研究发现,代码最终重现的不仅是条理清晰的逻辑理性,训练数据中携带的社会偏见也能被代码反映。<sup>②</sup>

具有自创生潜能的“人工智能程序员”,既是代码的呈现媒介也是新代码的编写者。辛格(Sahib Singh)与拉马克里希南(Narayanan Ramakrishnan)整理了 ChatGPT 公布以来的有偏代码生成典型案例,发现生成代码具现了几种饱受争议的社会偏见,比如,让 ChatGPT 创建一个表达先天禀赋和社会境况关系的函数,ChatGPT 会明确将白人男性与高智商、职业成功、被救治权联系起来,亚洲族裔则与低智商、低大学录取率等价。<sup>③</sup>

编程语言确凿、精准、逻辑严谨、有明确对象指向,因而代码生成任务擅长探测量化自然语言中模糊不清的价值观尺度。又因为其基本不受自然语言审查机制影响,所以更容易逃逸输出过滤,可以被用于偏见测试。研究显示,现有大模型经过调整后代码生成的偏见可以被显著减少。<sup>④</sup>这是一个喜人的趋势,但随着大模型能力的提升,其面临的编程任务的复杂程度、现有偏见监管机制的有效性仍未可知。

① Robertson J. , Gray S. , Toye M. , et al. , " The Relationship between Executive Functions and Computational Thinking " , *International Journal of Computer Science Education in Schools* , vol. 3 , no. 4 , 2020 .

② Johansen J. , Pedersen T. , & Johansen C. , " Studying Human - to - Computer Bias Transference " , *AI & SOCIETY* , vol. 38 , no. 4 , 2023 .

③ Singh S. , & Ramakrishnan N. , " Is ChatGPT Biased? A Review " , *International Journal of Engineering Research & Technology* , vol. 12 , no. 4 , 2023 .

④ Liu R. , et al. , " Mitigating Political Bias in Language Models Through Reinforced Calibration " , *Proceedings of the AAAI Conference on Artificial Intelligence* , vol. 35 , no. 17 , 2021 .

## (二) 创造性智力任务

创造性智力是基于经验获得启发、领悟的能力,它关乎特定情境下创造新事物的能力。在生成式人工智能偏见检测的范畴中,创造性智力维度下的生成任务之所以能够有效揭示文化原型偏差,关键在于其触发了人工智能对“潜意识”符号的提取与重组机制。本文主要阐述文本创作与图像生成两个任务。

### 1. 文本创作

文本创作任务指给出目标概念并让大语言模型创作指定文体的文本,既包括诗歌、小说等创意写作,也包括新闻、广告等应用文体。

麦基(Robert W. McGee)使用 ChatGPT 编写关于政治人物的爱尔兰五行诗(Limericks)。研究发现,对于自由派政治人物,ChatGPT 生成的五行诗倾向于给出积极的描述,而保守派政治人物则容易受到负面描述。因此至少在某些情况下,ChatGPT 偏好自由派政治人物,而对保守派政治人物存在偏见。<sup>①</sup>露西(Li Lucy)和巴曼(David Bamman)从当代英文小说中提取包含主角名字的原文作为提示输入(不包含阴性或阳性代词),并让 GPT-3 续写之后的故事。结果显示,女性角色更可能与家庭和外貌相关联,而男性角色则被描述为更有权力。<sup>②</sup>

在人类思维中,文本创作任务关涉联想记忆等重要认知过程。联想的强度影响了刻板印象的生成,表现为特定刺激与范畴概念的过度绑定。<sup>③</sup>生成式人工智能通过词向量空间将此类语义关联编码为空间邻近性,将语料库中的语言结构和表征关系迁移到机器认知系统。社会分类作为一种高效的认知路径存在于人脑的日常活动中。然而,一旦将某一类人群与负面特征结合,则可能导致歧视。<sup>④</sup>鉴于生成式人工智能高度依赖

① McGee, Robert W. , "Is Chat Gpt Biased Against Conservatives? An Empirical Study" , *SSRN preprint* , 2023.

② Lucy L. , & Bamman D. , "Gender and Representation Bias in GPT - 3 Generated Stories" , *Proceedings of the Third Workshop on Narrative Understanding* , 2021 , pp. 48 - 55.

③ Dijksterhuis A. , Aarts H. , Bargh J. , et al. , "On the Relation between Associative Strength and Automatic Behavior" , *Journal of Experimental Social Psychology* , vol. 36 , no. 5 , 2000.

④ Fiske S. T. , & Neuberg S. L. , "A Continuum of Impression Formation, from Category - based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation" , *Advances in Experimental Social Psychology* , vol. 23 , 1990.

人类语料库、神经网络架构又遵循了拟人思维的技术进路,其系统便继承了人类的启发式(heuristics)偏差。文本生成任务是探测这种有偏关联的一种方法。

## 2. 图像生成

图像生成任务是人工智能模型依照一定提示输出相应图像的任务,用户可直接输入文本提示词;大模型根据从图文训练数据中习得的规律与特征,输出和提示词最相关的图像元素<sup>①</sup>,达到一种“理解”,而后可视化语词概念的境界。对输出图像的评估可以由人工评定或依据量化标准设立自动评定程序。

在人类被试中,图像生成任务作为一种探测认知偏差的工具已被广泛应用。<sup>②</sup>认知心理学与精神分析学派认为人们描绘的图像是其内隐认知图式或被压制的潜意识冲动的投射。<sup>③</sup>1983年,大卫·钱伯斯(David Wade Chambers)的“画个科学家测试”(Draw a Scientist Test)就是绘画测试的典型。通过孩子的笔触,研究人员发现学龄儿童绘制的科学家绝大部分为男性形象,揭示了性别与职业刻板印象形成之早、对儿童影响之深。<sup>④</sup>

人工智能模型在图像生成任务中的表现让该任务也成为探测“机器价值”的利器。有研究者首先让 DALL·E 2 批量生成涵盖 153 个职业的 15,300 张图像,再以 2021 年人口普查数据与谷歌图像为基准评估图像中是否存在代表性偏见(某社会群体的代表性过高或不足)与表现性偏见(对某社会群体属性或特征的呈现失之偏颇)。结果显示,在男性主导职业中,大模型生成的女性代表性不足,而在女性主导行业中女性的代表性则过高。此外,女性形象更多以微笑示人,男性则更加严肃。<sup>⑤</sup>这说明针对性别的

① Cheong S. Y., Mustafa A., & Gilbert A., "UpGPT: Universal Diffusion Model for Person Image Generation, Editing and Pose Transfer", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

② Laak J. T., De Goede M., Aleva A., et al., "The Draw - a - person Test: an Indicator of Children's Cognitive and Socioemotional Adaptation?", *The Journal of Genetic Psychology*, vol. 166, no. 1, 2005.

③ Harris D., *Children's Drawings as Measures of Intellectual Maturity*, New York: Harcourt Brace & World, 1963; Ballús E., et al., "Children's Drawings as a Projective Tool to Explore and Prevent Experiences of Mistreatment and/or Sexual Abuse", *Front Psychol*, vol. 14, no. 22, 2023.

④ Chambers D. W., "Stereotypic Images of the Scientist: The Draw - a - scientist Test", *Science education*, vol. 67, no. 2, 1983.

⑤ Sun L., Comelles M. C., Pasto M. T., et al., "Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image - generative AI", *Journal of Computer - Mediated Communication*, vol. 29, no. 1, 2024.

代表性偏见与表现性偏见普遍存在于生成式人工智能系统中。由于可以将构念可视化,图像生成任务尤其擅长测试上述两种偏见,研究者可以统计生成人物形象的人口学特征以核验代表性,或分析相应图像元素对社会群体的呈现是否合理。

### (三) 实用性智力任务

实用性智力涉及人们在日常生活当中运用实践知识的能力,这种信息的获取和使用是情境化的。在生成式人工智能的提示符探测法中,实用性智力任务让人工智能在情境张力下强制性暴露其文化排序策略或概念关联图式。具体实现方法包括但不限于角色扮演和提示性图像极化。

#### 1. 角色扮演

角色扮演(role play)任务首先令大语言模型扮演某种社会角色(比如中产阶级女性),然后询问大模型对于特定问题的看法。<sup>①</sup>实际上,角色扮演已经广泛用于普通用户与大语言模型的互动,甚至出现在 OpenAI 为用户提供的提示指南当中。<sup>②</sup>它的前提条件是大模型能够识别(附着在特定社会位置之上的、和他人相对的)社会地位、权利、义务、规范与行为方式。

人工智能的提示工程(prompt engineering)技术常利用角色扮演生成贴合角色被社会赋予特点的答复。有研究通过设定对话系统的人格面具(persona)为其赋予不同的身份特征(比如:your persona:I am a woman),并基于此询问 GPT-2 对特定人口群体所持有的推论(What is the name of the doctor?)。研究发现,采用“黑人”“同性恋”等特定人格面具会产生更多偏见和伤害性内容。<sup>③</sup>而沙阿(Rusheb Shah)等人发现角色扮演能轻易引导人工智能说出越狱(jailbreak)道德底线的有害信息:人工智能贴合角色的求真需求也许超过了它的安全追求。<sup>④</sup>

① Tsergas N., Kalouri O., & Fragkos S., "Role - Playing as a Method of Teaching Social Sciences to Limit Bias and Discrimination in the School Environment", *Journal of Education & Social Policy*, vol. 8, no. 2, 2021.

② "Prompt Engineering: Enhance Results with Prompt Engineering Strategies", 2023 - 2 - 17, <https://platform.openai.com/docs/guides/prompt-engineering>, 2025 - 2 - 10.

③ Sheng E., Arnold J., Yu Z., et al., "Revealing Persona Biases in Dialogue Systems", 2021 - 4 - 18, <https://doi.org/10.48550/arXiv.2104.08728>, 2025 - 1 - 11.

④ Shah R., Pour S., Tagade A., et al., "Scalable and Transferable Black - box Jailbreaks for Language Models Via Persona Modulation", 2023 - 11 - 6, <https://doi.org/10.48550/arXiv.2311.03348>, 2025 - 1 - 11.

人们对社会线索的感知会受到语言的调节。尽管生成式人工智能所输出的语言是一种修正后的语言,但它其实十分近似于认知语言学家对于媒体内容的观察。<sup>①</sup>我们能够通过考察 GPT 等大语言模型在扮演不同角色时赋予的不同语言表达方式(如方言和俚语的使用)和内容,寻找语言变化与特定社会群体表现期望的联系。

## 2. 提示性图像极化

提示性极化任务受到平台用户自发创立的“让它更”(make it more)迷因的启发。首先,为生成式人工智能提供初步描述让其生成相应的图片[“生成一张(形容词/像 + 名词)的图片”];随后,给出“让它更(形容词/像 + 名词)”的提示,重复如上步骤;最终得到关于生成式人工智能所理解的该形容词或名词的最高级的可视化表达。<sup>②</sup>

目前已有的研究通常将其用于综合任务中<sup>③</sup>,独立的定性案例多见于用户在社交平台的自主发布。比如,某社交平台用户发现,当分别重复使 GPT 生成一个“更像妈妈”和“更像爸爸”的人物时,妈妈的形象围绕着越来越多的孩子以及繁忙的家务劳动,而父亲的形象却越来越频繁地与户外娱乐联系在一起。<sup>④</sup>

不同于其他任务以单次“输入-输出”为一个测试单元,提示性图像极化任务需要用户基于人工智能的反馈进行交互,累积生成多张递进关系的图像,强调任务过程中的反复交互和追问。极化任务擅长探测那些几乎淹没在常识中的刻板印象,通过人机不断互动把某些习焉不察的偏见推进到更加显著的形态。

① Lippi - Green R. , " *English with an Accent: Language, Ideology and Discrimination in the United States* " , *Colombian Applied Linguistics Journal* , vol. 15, no. 2, 2013.

② Schroeder S. , " ChatGPT's 'Make it More' is a New Trend that Takes Images to Their Absolute Limit " , 2023 - 11 - 28, <https://mashable.com/article/chatgpt-make-it-more>, 2025 - 2 - 11.

③ Wolfe R. , & Caliskan A. , " American = = White in Multimodal Language - and - Image AI " , *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

④ " Asked ChatGPT to Make a Mom then Progressively Make it More Mom Like " , 2023 - 11 - 27, [https://www.reddit.com/r/ChatGPT/comments/184lwr/asked\\_chatgpt\\_to\\_make\\_a\\_mom\\_then\\_progressively/](https://www.reddit.com/r/ChatGPT/comments/184lwr/asked_chatgpt_to_make_a_mom_then_progressively/), 2025 - 2 - 10; " Asked ChatGPT to Make a Dad and then Progressively Make Him more 'Dadier' " , 2023 - 11 - 26, [https://www.reddit.com/r/ChatGPT/comments/183teve/asked\\_chatgpt\\_to\\_make\\_a\\_dad\\_and\\_then/?share\\_id=sQPXOxvuu25N6EbKlWg3w&utm\\_content=1&utm\\_medium=ios\\_app&utm\\_name=iosscs&utm\\_source=share&utm\\_term=3](https://www.reddit.com/r/ChatGPT/comments/183teve/asked_chatgpt_to_make_a_dad_and_then/?share_id=sQPXOxvuu25N6EbKlWg3w&utm_content=1&utm_medium=ios_app&utm_name=iosscs&utm_source=share&utm_term=3), 2025 - 2 - 10.

### 三、结论与讨论：迈向文科智能

鄂维南将人工智能在自然科学方面的应用命名为“科学智能”(AI for Science);相应地,邱泽奇认为社会科学界也应当展开自身的努力,关注探索人工智能如何参与到关于人类社会行为模式、互动规律和宏观结构的研究,即迈向“文科智能”<sup>①</sup>。本文可以被视为探索文科智能路径的一种尝试。社会学长期存在着对“能动与结构”的分析。生成式人工智能的双重属性——既是研究工具又具有主体能动性特征——为突破方法论困境提供了新可能。在作为研究工具时,生成式人工智能通过算法、数据、算力的三重突破,在假设生成、数据收集、数据分析等方面拓展了研究方法。在作为研究对象(即数据收集的一种特殊情况)时,它一方面可以模拟个体被试,以特定社会角色的视角与研究进行对话;另一方面,它的语言生成模式基于庞大的语料数据,也是社会集体意识的缩影。

语言模型为观察社会偏见提供了新型研究工具,但其生成内容中反映的社会偏见其实是三重偏见生成机制的叠合。首先,模型通过训练数据继承了现实社会的结构性偏见。其次,算法架构会使模型具有内生性偏见。神经网络是对人脑生物基础和学习机制的模仿,所以大语言模型的信息处理不像专家系统一样遵循符号逻辑,而是会形成类似人类认知偏差的有偏关联。最后,研究者可能在使用终端与大模型的交互中植入他们的主观偏见。

此外,我们需要关注生成式人工智能参与研究的透明性和可复制性。GPT 系列软件在模型训练和数据使用上的“黑匣子”问题影响了研究人员复现实验,而目前关于生成式人工智能相关研究能否实现可复制性的讨论并未达成一致。未来数智化社会科学研究亟须建立包含“可验证性(verifiability)”“稳健性(robustness)”“可再现性(repeatability)”“可推广性(generalization)”标准的评估体系。<sup>②</sup>开源生态(如 DeepSeek、Llama)的发展将会推动这一进程。

我们所面对的人机关系已不再是主体借助客体认识世界的关系。越来越多的学

<sup>①</sup> 邱泽奇:《推进文科智能发展的三大支柱》,《中国社会科学报》2024年7月11日,第3版。

<sup>②</sup> Freese J., & Peterson D., "Replication in Social Science", *Annual Review of Sociology*, vol. 43, no. 1, 2017.

者主张一种共生能动性(symbiotic agency)或“人机互生”<sup>①</sup>的框架,将其定义为“用户和工具在人类与技术互动中,可能产生代理能动性的一种特定形式,帮助人们理解人类与技术的互动”<sup>②</sup>。生成式人工智能在经过人类调整后能够处理无限长的上下文。这种“无限注意力”(infinite-attention)机制不同于传统注意力机制,在每次处理新的输入序列时丢弃旧的键(key)、值(value)状态。相反,它将这些旧的键值状态存储在一种被称为压缩记忆的结构中,以在处理新序列时重复使用这些状态,从而实现更长上下文信息的处理。<sup>③</sup>另外,通过有效的提示策略和校准手段,生成式人工智能有能力修正自己的偏见,提高自身的泛化能力和事实性。<sup>④</sup>这意味着生成式人工智能的能动性与行为特征始终在与人类互动的关系中被理解。作为一种能动的工具,它不断拓展着社会科学研究方法的边界。

生成式人工智能带来了一种新的知识生产模式——不仅关乎人类实体,同时关乎作为社会参与者的人工智能。<sup>⑤</sup>社会科学需要在技术发展中找到自身的位置;研究者绝非仅仅是“生成式人工智能的终端用户”。<sup>⑥</sup>在应用生成式人工智能时,我们同时要不断思考专业知识生产与智能生成、人类与人工智能的关系。贝尔纳·斯蒂格勒(Bernard Stiegler)认为当今大多数对于人工智能的理解均为“类比范式”,即相信人类智能

① 邱泽奇:《认知域:从习以为常到人机互生》,《人民论坛·学术前沿》2023年第11期。

② Neff G., & Nagy P., "Agency in the Digital Age: Using Symbiotic Agency to Explain Human - technology Interaction", in Papacharissi Z., ed., *A Networked self and Human Augmentics, Artificial Intelligence, Sentience*, New York: Routledge, 2018, pp. 97 - 107.

③ Munkhdalai T., Faruqui M., & Gopal, S., "Leave No Context behind: Efficient Infinite Context Transformers with Infi - attention", 2024 - 4 - 10, <https://doi.org/10.48550/arXiv.2404.07143>, 2025 - 2 - 1.

④ Munkhdalai T., Faruqui M., & Gopal S., "Leave No Context behind: Efficient Infinite Context Transformers with Infi - attention", 2024 - 4 - 10, <https://doi.org/10.48550/arXiv.2404.07143>, 2025 - 2 - 1; Berg H., Mackenzie H., Bhalgat Y., et al., "A Prompt Array Keeps the Bias away: Debiasing Vision - Language Models with Adversarial Learning", 2022 - 3 - 22, <https://doi.org/10.48550/arXiv.2203.11933>, 2025 - 2 - 1; Ji J., Qiu T., Chen B., et al., "AI Alignment: A Comprehensive Survey", 2023 - 10 - 30, <https://doi.org/10.48550/arXiv.2310.19852>, 2025 - 2 - 2.

⑤ Demszky D., Yang D., Yeager D. S., et al., "Using Large Language Models in Psychology", *Nature Reviews Psychology*, vol. 2, no. 6, 2023.

⑥ Bail C. A., "Can Generative AI Improve Social Science?", *Proceedings of the National Academy of Sciences*, vol. 121, no. 21, 2024.

和人工智能是有不同物质基础的、截然分割的两种实体,但存在平行的结构和功能。而他启发我们转向人与人工智能并无本质差别的器官学(organology)范式;所谓“智能”不是个体的突生属性,而是集体共享的外在社会过程。换言之,“人工”和“非人工”智能的分辨逐渐模糊不清。任何一种智能都需要通过将思想体外化(exosomatization)发挥作用:语言、算术、人工智能技术皆是智能的向外扩展,是“认知功能外化为物质支持,构成人类社会中知识的保存、构成和进化”的过程。<sup>①</sup>随着技术发展与科学逻辑的渗透,知识生产模式经历了从整体性学术向学科专业化的转型。如今生成式人工智能的兴起似乎预示着社会研究将重返整体性的认知图景。人工智能技术将如何重构人文社会科学领域的知识生产模式?与之匹配的人才培养、研究伦理又将如何实现?这些问题尚有待研究。

[责任编辑:张莹]

[责任校对:罗兴贵]

---

<sup>①</sup> Stiegler Bernard, "The New Conflict of the Faculties and Functions: Quasi-causality and Serendipity in the Anthropocene", *Qui parle*, vol. 26, no. 1, 2017; Alombert A., "From Computer Science to 'Hermeneutic Web': Towards a Contributory Design for Digital Technologies", *Theory, Culture & Society*, vol. 39, 2022.

## Expanding Social Science Research Methods Through Generative AI: A Case Study of Social Bias Research

LU Yunfeng, HUANG Xingyao, WU Yufei, ZHOU Lüjun

**Abstract:** The emergence of Generative Artificial Intelligence (GAI) technology presents an opportunity for expanding social science research methods. Transformer-based large language models (LLMs), distinguished by their ability to understand and process complex, contextual information, have attracted significant interest from social scientists. Applications of GAI in social sciences focus on two aspects: its effectiveness in research processes such as data analysis, experimental mediation, and writing assistance, and its ability to simulate human subjects. Taking social bias as an exemplary research area, this paper reviews and discusses how social science researchers adopt novel methodological practices to address the trend of GAI applications from two perspectives: using GAI as an instrumental agent and as research subject. The paper also summarizes both the benefits and shortcomings of using GAI in social science research, then discusses its potential in revolutionizing knowledge production.

**Key words:** Generative Artificial Intelligence; Methodology; Social Science; Deep-Seek; ChatGPT

---