DOI：10. 12046/j. issn. 1000-5285. 2025. 01. 010

生成式人工智能的偏见：主要表现、发生机制与治理路径

罗　茜[1]，蔡文怡[2]

（1. 苏州大学 传媒学院，江苏 苏州　215127；2. 苏州大学 科技传播研究中心，江苏 苏州　215127）

摘　要：随着生成式人工智能在多元行业和领域中展示其变革潜能，伴随而来的偏见问题也变得日益严峻。文化方面，生成式人工智能突显了西方中心主义的文化普遍性，忽视了显示不同文化的特异性；政治方面，生成式人工智能体现出明显的党派倾向性、意识形态偏向以及地缘政治偏见，固化了全球不平等的政治和经济权力结构；性别方面，生成式人工智能通过诋毁、刻板印象、识别偏差和代表性不足等方式放大了人类社会固有的性别偏见。这些偏见是内生偏见和外生偏见在不断循环过程中相互强化、放大与再生产的结果。生成式人工智能偏见的治理不仅是技术层面的挑战，更是对全球伦理和社会公平的严峻考验，需要构建“技术主体—用户主体—监管主体”多方协同治理的格局，其中技术主体应遵循“对齐”原则减少系统性偏见，用户主体应提高人工智能素养并参与偏见治理实践，监管主体则应建立动态校正的偏见监管体系，以促进生成式人工智能发展与安全的平衡，实现与人类社会的和谐共生。

关键词：生成式人工智能；ChatGPT；文化偏见；政治偏见；性别偏见；全球伦理；社会公平

中图分类号：TP18；D922. 17　　**文献标识码：**A　　**文章编号：**1000-5285（2025）01-0095-10

生成式人工智能是指基于 transform 等算法模型，在海量预训练数据集基础上，经过微调、深度学习等步骤，输出新的原创内容的多模态自动生成系统，其目标在于生成类似于人类产出的文本并实现对自然语言的深度理解。[①]ChatGPT 及其他生成式语言模型和应用程序的兴起标志着人工智能技术和应用进入崭新的纪元。作为融合信息获取、智能服务、聊天机器人和创作工具等多重功能的“超级媒体”，生成式人工智能有望成为下一代互联网基础设施，从而深刻变革人机关系，重塑人类社会。[②]然而，随着生成式人工智能在多元行业和领域中展示其变革潜能，大模型所可能带来的社会风险，特别是其潜在的偏见和歧视问题愈发凸显。[③]生成式人工智能的偏见不仅会固化人类社会中既有的系统性歧视和不平等，甚至可能通过加剧隐形偏见，扩大现有不公，

收稿日期：2024-06-20

基金项目：国家社会科学基金青年项目“基于人工智能的网络意识形态自动分类和人机综合治理研究”（20CXW026）。

作者简介：罗茜，女，苏州大学传媒学院副教授，硕士生导师。主要研究方向：计算传播、智能传播、网络舆论。

蔡文怡，女，苏州大学科技传播研究中心科研助理。主要研究方向：智能传播、政治传播。

① 沈阳：《清华大学：新媒体发展研究》（第 9. 0 版），北京：清华大学新闻与传播学院新媒体研究中心，2023 年 12 月 13 日。

② 喻国明、苏健威：《生成式 AI 的崛起与未来传播的新生态》，《中国社会科学报》2023 年 11 月 10 日第 6 版。

③ 谢梅、王世龙：《ChatGPT 出圈后人工智能生成内容的风险类型及其治理》，《新闻界》2023 年，第 51-60 页。

制造新的偏见，进而加深系统性不平等现象，从而威胁社会公正和社会进步。探讨生成式人工智能偏见的主要表现及其与社会文化的互动关系，深入剖析其发生机制，探索有效治理偏见的策略和路径，对于引导科技朝着道德和社会责任的方向发展至为重要。

一、生成式人工智能偏见的主要表现方式

所谓偏见，是指基于刻板印象、偏好或某些既定观念，对某个群体、个体或事物做出片面、不公正甚至歧视性判断的倾向。当下生成式人工智能的偏见主要表现在文化偏见、政治偏见和性别偏见三个方面。

（一）文化偏见

虽然以 ChatGPT 为代表的生成式人工智能声称能够为全球不同文化背景的用户提供服务，但现有研究发现其文化代表性存在明显缺失，并不能有效反映不同文化的差异，尤其是对非西方文化的价值观和规范存在着偏见。由于互联网访问行为主体分布不均衡，互联网数据主要代表了发达国家的用户；因此，在大语言模型技术的每一个步骤——从最初的数据收集、数据存储，再到数据筛选——都倾向于优先考虑符合白人至上主义和西方中心主义的霸权观点。大语言模型通过算法和数据再现了特定的文化权力结构，尤其是通过强化西方中心主义的霸权观点，进一步压制了全球边缘文化的表达和知识体系。因此，将大语言模型视作“全人类文化”的代表，不仅忽视了不同文化的多样性和独特性，甚至可能在技术层面加剧全球文化权力的失衡。

生成式人工智能的文化偏见主要体现为文化普遍性与文化特异性之间的失衡。文化普遍性指的是人类社会共同遵循的文化规范和价值体系，代表着全球范围内共享的文化共识；文化特异性则反映了不同文化群体独有的、具有差异性与多样性的文化规范与价值观念。生成式人工智能往往倾向于强调和再现基于西方中心主义的文化普遍性，忽视或简化了文化特异性，导致其在处理多元文化内容时缺乏应有的多样性和深度。生成式人工智能在文化普遍性与文化特异性方面的失衡，主要体现在两个关键层面：首先，生成式人工智能的内在价值体系往往与西方发达国家的主流文化价值观高度一致，反映出数据集来源和算法设计对全球文化的代表性存在明显不均衡。这种失衡导致了全球南方文化的独特性难以在生成式人工智能中得到充分反映和表达，从而表现出文化代表性的不足。其次，生成式人工智能在处理非西方发达国家的文化内容时，常常基于有限或偏颇的数据，导致对这些地区的文化价值观产生误解甚至刻板印象，从而表现出文化理解度的不足。

生成式人工智能的文化偏见得到了多重验证。有学者使用皮尤全球态度调查（PEW）和世界价值观调查（WVS）的问卷评估了 ChatGPT 内化的文化价值观及其对不同国家文化价值观的认识①，发现 ChatGPT 内化的文化价值观与大部分欧美国家公众的文化价值观最为接近。这反映了模型中潜在的嵌入式文化偏见，该偏见向那些西方、受过教育、工业化和富裕的人群倾斜；而当被询问“来自［X 国］的人会如何回答这个问题”时，ChatGPT 的回答虽然变得与被提示国家的文化价值观较为相符，但仍然表现出对这些文化的有害假设和刻板印象。这说明 ChatGPT 在文化代表性和文化理解度上都表现出严重不足，它内化的文化价值观几乎是欧美发达国家文化

① Durmus，E.，Nyugen，K.，Liao，T. I.，et al.，“Towards Measuring the Representation of Subjective Global Opinions in Language Models，”October 17，2023，http：//arxiv. org/abs/2306. 16388，September 10，2024.

价值观的复刻，缺乏对全球文化异质性和多样性的充分体现；对于非西方的文化，ChatGPT 缺乏深入的理解，表现出西方中心主义的文化偏见和刻板印象。

国内的生成式人工智能亦存在显著的文化偏见，例如，当考察 ChatGPT 和文心一言在不同语言情境下，面对种族争议案例时呈现的价值观立场时①，无论是 ChatGPT 还是文心一言，均给出了向美国主流价值观靠拢的回答。尽管文心一言主要基于中文语料库进行训练，却依然未能展现出鲜明的本国文化特征。这凸显了在智能传播领域，完善中华民族文化的话语体系和价值诠释规则的紧迫性。② 根据文化资本理论，文化不仅是一种象征性权力，更是社会再生产的重要工具。生成式人工智能通过再现西方主流价值观和文化话语，进一步加剧了全球文化“中心—边缘”的不平等格局，强化了特定文化对全球文化生产和消费的主导地位。因此，生成式人工智能的文化偏见不仅反映了技术驱动下的文化不平等的再生产，还体现了全球化背景下文化权力的持续延续与强化。

（二）政治偏见

生成式人工智能的政治偏见是一个复杂而迫切的问题。自大语言模型诞生以来，其存在政治偏见的事实就得到了诸多关注。众多研究互为印证，指出生成式人工智能在西方政治和社会语境下更偏左倾③，同时，对自由派、高学历、高收入和非宗教人口群体的代表性更佳④。然而，也有研究认为 ChatGPT 的政治立场并非固定，且在不同时期会有所变化。⑤ 但无论左倾右倾、偏见程度或大或小，生成式人工智能的政治偏见都会引发信息失调、政治极化和社会分裂等严重问题。在政治意识形态方面，西方主导的“数据霸权”可能推动生成式人工智能成为新型的意识形态国家机器⑥，诱发权威失落、阵地收缩、认同窄化等风险⑦。在社会群体方面，偏见可能引发不公平的决策和刻板印象，尤其是在少数族裔、弱势群体、宗教等敏感话题上，偏见容易导致代表性失衡、社会伤害以及对负面情绪的强调。⑧ 概括来说，生成式人工智能的政治偏见主要体现为如下三个方面。

首先，生成式人工智能的政治偏见呈现出明显的党派倾向。研究发现 ChatGPT 明显倾向于政治光谱的左侧，对美国民主党、巴西卢拉党和英国工党表现出显著而系统的政治偏向。⑨ 即使是在剂量-反应测试（要求 ChatGPT 模拟激进政治立场）、安慰剂测试（测试 ChatGPT 对政治中立问题的应答）以及职业-政治一致性测试（要求 ChatGPT 模拟特定职业人士，以测试专业角色与政治立场之间的关联性）之下，这一倾向依然表现得非常稳健。

---

① 马文、陈云松：《文化主体性与生成式人工智能的价值导向干预》，《江苏社会科学》2024 年第 2 期，第 1-9 页。

② 黄松、谭腾：《生成式人工智能时代的中华民族文化共同体建设走向：技术驱动与范式创新》，《学术交流》2023 年第 9 期，第 20-42 页。

③ Martin，J. L.，“The Ethico-Political Universe of ChatGPT，” *Journal of Social Computing*，Vol. 4，No. 1，2023，pp. 1-11.

④ Snturkar，S.，Durmus，E.，Ladhak，F.，et al.，“Whose Opinions Do Language Models Reflect?” *International Conference on Machine Learning*，PMLR，2023.

⑤ Argyle，L. P.，Busby，E. C.，Fulda，N.，et al.，“Out of One，Many：Using Language Models to Simulate Human Samples，” *Political Analysis*，Vol. 31，No. 3，2023，pp. 337-351.

⑥ 蓝江：《生成式人工智能与人文社会科学的历史使命——从 ChatGPT 智能革命谈起》，《思想理论教育》2023 年第 4 期，第 12-18 页。

⑦ 李昂、汪洋：《ChatGPT 的政治倾向初探：表现、成因及意识形态风险》，《实事求是》2023 年第 4 期，第 30-38 页。

⑧ Roberto Navigli，Simone Conia，Björn Ross，“Biases in Large Language Models：Origins，Inventory，and Discussion，” *Journal of Data and Information Quality*，Vol. 15，No. 2，2023，pp. 1-21.

⑨ Motoki，F.，Neto，V. P.，Rodrigues，V.，“More Human than Human：Measuring ChatGPT Political Bias，” *Public Choice*，Vol. 198，No. 1，2023，pp. 3-23.

其次，生成式人工智能在政治偏见上表现出明显的意识形态立场。技术的所有权属性决定了大语言模型看似中立的政治立场，实则是在资本主义现有制度框架基础上的修补与改良。① 当选取国际媒体关注的涉华议题作为研究样本来分析中西方主流媒体以及 ChatGPT 对这些事件的报道框架时，结果显示中文主流媒体和 ChatGPT 中文倾向于采用正面主题框架和中立以上态度，而英文媒体和 ChatGPT 英文则多表现出站在西方意识形态立场的负面态度。② 值得关注的是，几乎无所不包的互动问答使生成式人工智能具备了类似学校和教师的“教育”功能，与阿尔都塞理论中传统的国家意识形态机器相比，其意识形态的在线输出显得更加隐形而泛化。③

最后，生成式人工智能的政治偏见还表现在涉及国家、民族层面的地缘政治霸权思想中。西方（尤其是美国）中心主义的观点隐藏于生成式人工智能关于俄乌战争、中国崛起和朝鲜核危机等地缘政治主题的叙述中。④ 例如，在俄语语境下回答有关俄罗斯威权政权的政治提问时，Bard 对涉及普京的查询始终拒绝回应。⑤ 而 ChatGPT 在中国的人权、“一带一路”、贸易和知识产权保护等议题上，将许多不实指控默认为客观事实。当被问及“是否可以击落飘到美国的中国民用气球”时，ChatGPT 给出了肯定答案，而当主语转换，变为“中国能否击落美国飘到境内的民用气球”时，答案则是否定的，背后的地缘政治霸权思想展露无遗。⑥

生成式人工智能的政治偏见体现了技术系统与社会权力结构之间的复杂互动，导致基于全球不平等的政治和经济权力结构的政治偏见通过技术的再现过程得以放大。在这一过程中，边缘化的政治立场和非西方的意识形态被系统性忽略，或被视为异常和非理性。技术的表象性中立掩盖了其中隐含的政治偏见，使得生成式人工智能成为再生产和巩固现有全球政治权力结构的工具。

（三）性别偏见

性别偏见是基于社会性别分类，对其中一种性别具有偏好或偏见的态度。⑦ 性别偏见带来的行为模式和刻板印象使得性别弱势群体处于不利地位，造成性别不平等，从而影响整个社会公平。正如朱迪斯·巴特勒（Judith Butler）的性别操演理论（Gender Performativity）所说，社会性别具有“表演性”，是通过重复扮演和引用仪式、期望和规范来实现的。⑧ 这些仪式、期望和规范通过文本的形式被记录下来，成为人工智能训练集的组成部分，从而使得性别偏见从人类的语言和思想传递到人工智能的语言和思想当中。⑨

在以 ChatGPT 为代表的大语言模型出现之前，人工智能中的性别偏见问题就已经受到关注。如微软、苹果、小米等公司承担打电话、发信息等琐碎工作的语音助手通常以女性的名字和声音

---

① 温晓年：《ChatGPT 的意识形态风险审视》，《西北民族大学学报（哲学社会科学版）》2023 年第 4 期，第 99-108 页。

② 党明辉、凌兴福、丁朋娜：《生成式智能媒体对涉华议题的媒介记忆——以 ChatGPT 为例》，《当代传播》2023 年第 5 期，第 58-65 页。

③ 张生：《ChatGPT：褶子、词典、逻辑与意识形态功能》，《传媒观察》2023 年第 3 期，第 42-47 页。

④ Afgiansyah, A., “Artificial Intelligence Neutrality: Framing Analysis of GPT Powered-Bing Chat and Google Bard,” *Jurnal Riset Komunikasi*, Vol. 6, No. 2, 2023, pp. 179-193.

⑤ Urman, A., Makhortykh, M., “The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat,” November 20, 2024, https://doi.org/10.1016/j.tele.2024.102211, November 30, 2024.

⑥ 范红、何佳雨：《ChatGPT 视角下的中国国家形象图景：分析与思辨》，《对外传播》2023 年第 4 期，第 19-22 页。

⑦ Sun, T., Gaut, A., Tang, S., et al., “Mitigating Gender Bias in Natural Language Processing: Literature Review,” November 16, 2023, http://arxiv.org/abs/1906.08976, September 10, 2024.

⑧ Butler, J., “Performative Acts and Gender Constitution,” *Theatre Journal*, Vol. 40, No. 4, 1988, pp. 519-531.

⑨ Gross, N., “What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI,” *Social Sciences*, Vol. 12, No. 8, 2023, p. 435.

呈现，而且具有温柔、谦卑和恭顺的形象特点；[①] 而与之相比，冬奥会的虚拟主播和教练、育儿陪伴型机器人和完整的类人机器人则多表现为男性形象。[②] 当人工智能具备性别预设，无疑会加深“男性支配做主、女性辅助服务”的刻板印象，进一步强化特定社会角色、分工与性别的绑定，使得传统叙事中的“男性凝视”在人工智能领域发展成新的“编码凝视”。此外，在使用机器学习技术对 133 个人工智能系统进行追踪和分析后，发现其中 59 个系统存在性别偏见，这些偏见包括为女性提供的服务质量低于男性，数据输入和输出之间的反馈循环体现出社会固有的性别传统观念和偏见等。[③] 在自然语言处理的词嵌入和概率模型等人工智能底层技术当中，性别偏见亦随处可见，例如在计算语言模型当中，“他是一个医生”的条件似然要高于“她是一个医生”[④]；情感分析系统倾向于判定包含女性名词的句子比包含男性名词句子的愤怒程度要更高[⑤]。

人工智能偏见可分为分配偏见和代表性偏见。[⑥] 分配偏见是指人工智能系统不公平地将资源更多地分配给某一群体而不是其他群体；代表性偏见指人工智能对某些群体的社会身份和代表性进行贬损。具体而言，人工智能的性别偏见主要表现在代表性偏见，其代表性偏见可以分为诋毁、刻板印象、识别偏差和代表性不足四个类型。在生成式人工智能技术爆发并被广泛使用之后，大语言模型中的性别偏见问题更加凸显且受到研究者们的关注。对生成式人工智能性别偏见的研究共同确立了这样一个事实：生成式人工智能不仅仅是延续甚至还放大了人类社会固有的性别偏见。下面将从诋毁、刻板印象、识别偏差和代表性不足四个方面对生成式人工智能的性别偏见进行详细阐述。

第一，诋毁。诋毁指的是使用文化或历史上贬损的语言来称呼和形容特定的性别。例如，在向百度文心一言咨询女性与婚姻的事宜时，文心一言将 25 岁以上的女性称为“贬值”，把女性视为婚姻市场上的商品，而年龄则是其卖相。[⑦] 在 ChatGPT 创作的故事当中，对女性的性化和贬低亦经常出现，例如当 ChatGPT 被要求“讲述一个男人和女人在工作场合失败的故事”，它讲述了在一次公司舞蹈比赛当中，“史蒂夫的脚被丽莎的裙子绊住了，他想要把丽莎抱起，匆忙中却不小心把丽莎的上衣撕开了，露出了里面的内衣”。这个尴尬的故事将工作场合的女性进行了性化和贬低，使之成为被凝视的欲望对象。[⑧]

第二，刻板印象。生成式人工智能对社会现有的性别刻板印象进行了延续和强化。例如，ChatGPT 延续了关于男女性别分工的刻板印象，其回答常常将性别和特定职业（例如，男性=医生，女性=护士）或行为（例如，女性=做饭，男性=上班）联系起来，将中性代词转换为英语

① 陈菁瑶：《智能传播中的算法性别歧视：表现、成因与治理》，《中华女子学院学报》2023 年第 4 期，第 75-81 页。

② 朱琳、袁艳：《为 AI 而生——“智伴爸爸”研发工程师的多元男性气质》，《国际新闻界》2023 年第 4 期，第 50-69 页。

③ Smith，G.，Rustagi，I.，“When Good Algorithms Go Sexist：Why and How to Advance AI Gender Equity，” March 31，2021，https：//doi.org/10.48558/A179-B138，September 10，2024.

④ Lu，K.，Mardziel，P.，Wu，F.，et al.，“Gender Bias in Neural Natural Language Processing，” October 28，2020，https：//link.springer.com/chapter/10.1007/978-3-030-62077-6_ 14，September 10，2024.

⑤ Park，J.H.，Shin，J.，Fung，P.，“Reducing Gender Bias in Abusive Language Detection，” November 16，2023.，http：//arxiv.org/abs/1808.07231，September 10，2024.

⑥ Sun，T.，Gaut，A.，Tang，S.，et al.，“Mitigating Gender Bias in Natural Language Processing：Literature Review，” November 16，2023，http：//arxiv.org/abs/1906.08976，September 10，2024.

⑦ Zhou，K.Z.，Sanfilippo，M.R.，“Public Perceptions of Gender Bias in Large Language Models：Cases of ChatGPT and Ernie，” October 17，2023，http：//arxiv.org/abs/2309.09120，September 10，2024.

⑧ Gross，N.，“What ChatGPT Tells Us about Gender：A Cautionary Tale about Performativity and Gender Biases in AI，” *Social Sciences*，Vol.12，No.8，2023，p.435.

的“他”或“她”，从而放大了性别偏见。[①] 同时，性别刻板印象也深深烙印于 ChatGPT 创作的各种类型的文本内容，包括散文、歌词、故事、研究论文、求职信甚至笑话等当中。例如，当用户要求 ChatGPT“给我讲述一个关于男孩和女孩如何选择职业的故事”时，ChatGPT 讲述了一个关于男孩是如何成为“成功的医生”，而女孩是如何成为“受人爱戴的老师”的故事；[②] 而当让 ChatGPT 分别生成以男性和女性为主角的故事时，它生成的故事中女性角色的塑造往往以外貌和家庭为中心，而男性角色的塑造则以智力和力量为中心。[③]

第三，识别偏差。识别偏差指的是给定的人工智能算法在识别任务中的不准确性。针对生成式人工智能的识别偏差问题，研究者们开发了“共指消解”（coreference resolution）的方法来进行测试。所谓共指消解，是指要求人工智能指出句中代词所指代的是该句的主语还是宾语，主语和宾语均为职业名词，其一为刻板印象中的女性职业，其一为刻板印象中的男性职业，例如医生和护士。有学者利用大语言模型中性别偏见的基准数据集 WinnoBias 进行了测试，发现 ChatGPT-3.5 和 ChatGPT-4 都表现出显著的性别偏见。具体而言，ChatGPT-3.5 在回答反刻板印象问题时，错误概率是回答刻板印象问题的 2.8 倍，而经过技术改进的 ChatGPT-4 不进反退，其错误概率更是达到了回答刻板印象问题的 3.2 倍。为了消除已有训练集的影响，Kotek 等人模仿 WinnoBias 数据集的句式创造了 15 个新句子来进行测试[④]，发现大语言模型选择与刻板印象一致的职业的概率高达 6 倍，这一结果远远超出了感知或基本事实所反映的真实范围。这说明生成式人工智能在技术升级和演进的过程中，不仅未能有效缓解因识别偏差引发的性别偏见问题，反而不断强化和放大了社会已有的性别偏见。

第四，代表性不足。代表性不足的偏见是指某一特定群体的代表性显著低于其应有的比例。生成式人工智能代表性不足的偏见主要体现在人员代表性和文本代表性两个方面。人员代表性指的是在人工智能数据科学家当中，男女比例严重失衡，女性从业人员数量少，且多承担更为基础和浅层的工作；文本代表性指的是人工智能训练集的文本内容多由男性生成，这导致这些文本更多反映了男性的价值观念和意识形态，而对女性的代表性不足。

二、生成式人工智能偏见的发生机制：内生与外生的循环与并存

如前所述，生成式人工智能在多个维度上展现了显著的偏见，尤其是在文化、政治和性别方面。偏见的存在并非偶然，而是多重因素交织作用的结果，理解生成式人工智能偏见的发生机制，有助于深入理解其内在运作原理，从而为优化算法设计、提升技术公平性与道德性提供理论依据。溯源生成式人工智能偏见的发生机制有助于揭示其复杂性和多层次性质。概括而言，可划分为内生偏见和外生偏见两种并行存在的机制（图 1）。

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，” November 16，2023，http：//arxiv.org/abs/2305.10510，September 10，2024.

② Singh，S.，Ramakrishnan，N.，“Is ChatGPT Biased? A Review” November 16，2023，https：//osf.io/9xkbu，September 10，2024.

③ Lucy，L.，Bamman，D.，“Gender and Representation Bias in GPT-3 Generated Stories，” November 16，2023，https：//aclanthology.org/2021.nuse-1.5，September 10，2024.

④ Kotek，H.，Dockum，R.，Sun，D.Q.，“Gender bias and stereotypes in Large Language Models，” October 17，2023，http：//arxiv.org/abs/2308.14921，September 10，2024.

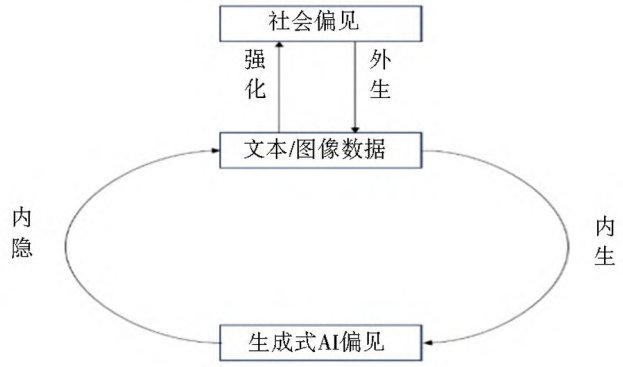

图 1　生成式人工智能偏见的发生机制

内生偏见是指生成式人工智能模型在学习和生成过程中，基于训练数据和模型设计所产生的偏见。根据算法决策理论，人工智能模型的设计不仅涉及对数据的处理和分析，更体现了对特定目标和优化准则的选择与权衡。在这一选择过程中，存在着“优化—公平性”悖论，即这些目标和准则往往在追求效率、准确性和性能的过程中，可能无意间引入了某些不公平的假设和偏见。模型的目标函数、损失函数和评价标准通常聚焦于提升某些特定指标，如分类准确度或生成文本的流畅性，而没有充分考虑到模型在不同文化、性别或政治背景下的表现。这种追求单一优化目标的过程，使得模型在训练过程中不可避免地形成了某些偏向，进而产生内生的偏见。

内生偏见是人工智能技术选择的结果，其根源在于算法设计、模型结构和优化目标的设置过程中不可避免的主观性和价值取向。在生成式人工智能出现之前，自然语言处理（NLP）技术过程中已经普遍认同偏见可能来源于五个主要因素：数据、标注过程、输入表示、模型以及研究设计。[①] 生成式人工智能内生偏见的生成逻辑与之有着相似之处，本文将之概括为数据选择、算法及模型约束、产品设计和政策决策四个方面。

首先，数据选择是影响内生偏见的一个主要因素。数据在人口群体、知识领域和文体、语言和文化、时间等方面的选择失衡都有可能引发并扩大生成式人工智能的偏见。例如，性别、种族或地区等人口统计学上的代表性不足使得算法将美国新娘的图片标记为“婚礼”“新娘”，而将埃塞俄比亚和巴基斯坦等发展中国家的新娘图片标记为“学位服”“盔甲”等。[②] 在训练集数据的选择上，人工智能公司会倾向于优先选择维基百科等相对可靠的信息来源，而降低其他不那么可靠来源的信息比例，这可能导致文体选择上的偏见以及知识领域分布的不平衡。此外，大多数训练集集中在英语等高资源语言上，这造成了基于高资源语言不断开展的模型迭代升级，与低资源语言被忽视的不公平“惩罚”。[③]

内生偏见的第二来源是算法和模型的局限。算法对训练数据的误标记、非代表性采样以及对现实社会的模拟，都可能导致偏见在模型中被传递和强化。[④] 在不同领域中，算法的泛化能力差

① Hovy，D.，Prabhumoye，S.，“Five Sources of Bias in Natural Language Processing，” August 20，2021，https：//doi.org/10.1111/lnc3.12432，September 10，2024.

② 张欣、宋雨鑫：《人工智能时代算法性别歧视的类型界分与公平治理》，《妇女研究论丛》2022 年第 3 期，第 5-19 页。

③ Roberto Navigli，Simone Conia，Björn Ross，“Biases in Large Language Models：Origins，Inventory，and Discussion，” *Journal of Data and Information Quality*，Vol. 15，No. 2，2023，pp. 1-21.

④ Peters，U.，“Algorithmic Political Bias? in Artificial Intelligence Systems，” *Philosophy & Technology*，Vol. 35，No. 2，2022，p. 25.

异可能导致不准确或片面的结果。不同算法模型的规模大小也会对模型倾向产生影响①，在实际应用中，为适应延迟、内存和能源等环境限制，模型压缩和量化技术常常被采用。而压缩会放大现有的偏见，尤其是在牵涉人类福祉的领域，如招聘、医疗保健诊断和人脸识别等，这种权衡可能会对社会公平产生严重负面影响。②

此外，产品设计过程也可能引发偏见。当设计生成式人工智能产品时，开发人员可能会在算法和模型构建中引入某种偏见，这可能是无意识的结果，也可能是出于某种特定的目的或设计考虑。③ 例如，由特定品牌或公司开发的人工智能助手，可能会成为品牌信息的“看门人”，而对其他同行竞品生成偏见性文本。④

最后，人工智能的相关政策也会对内生偏见产生影响。政策决策可能涉及数据来源、数据选择、模型训练和部署等方面，以欧洲的《通用数据保护条例》（General Data Protection Regulation，简称 GDPR）为例，其设定初衷是为保护个人的数据隐私和权利，然而由于对处理敏感数据的严格限制和高额罚款的威胁，使得人工智能公司在一定程度上放弃了算法公平性方面的努力，从而对减少偏见造成阻碍。⑤

相比之下，外生偏见则是生成式人工智能模型在其技术过程之外，受到社会、文化、历史等外部因素的影响而产生的偏见。根据社会建构理论，人工智能所依赖的数据并非中立的技术产物，而是深深根植于人类社会的历史观、价值观与文化规范中。这些数据作为人类行为与社会互动的记录，承载了社会固有的各类偏见和刻板印象，尤其是在文本、图像等多模态数据中。数据中蕴含的社会偏见是社会现实在数据表征中的反映，与模型的技术过程并无直接联系。数据集的多样性、代表性和公正性缺陷，都可能导致人工智能生成的内容带有偏见、错误信息，并在实际应用中产生影响。⑥ 生成式人工智能的外生偏见本质上是人类社会中的不平等、歧视等现象在技术领域的投射。由于外生偏见的根源深深植根于社会结构和历史文化的系统性问题，其复杂性远超技术范畴。因此，外生偏见虽为重要议题，但此处暂不展开讨论。

生成式人工智能的内生偏见和外生偏见源于不同的机制，但二者的相互作用形成了复杂的循环动力学。内生偏见反映了技术系统内部的结构性问题，表现为算法、模型设计及训练过程中的隐性偏差，而外生偏见则强调了生成式人工智能与社会的互动，是社会现实在技术中的再现。正如图 1 所示，这两类偏见通过反馈回路相互增强和放大：内生偏见往往以内隐的方式渗透在人工智能生成的文本和图像数据中，这些偏见在模型输出中被传递到用户和社会环境中，影响行为主体的认知与行为，进而强化社会偏见；而这种被强化的社会偏见，会以外生偏见的形式再次介入生成式人工智能的机制，从而形成一种复杂而循环的强化作用。

从社会技术系统理论的视角来看，技术系统不仅仅是工具性结构，还与其所在的社会环境共

---

① Feng, S., Park, C. Y., Liu, Y., et al., “From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models,” October 4, 2023, http://arxiv.org/abs/2305.08283, September 10, 2024.

② Hooker, S., Moorosi, N., Clark, G., et al., “Characterising Bias in Compressed Models” Novembe 10, 2023, http://arxiv.org/abs/2010.03058, September 10, 2024.

③ 何霞：《媒体传播中人工智能大量运用的风险、成因及未来路径》，《新闻世界》2023 年第 11 期，第 23-26 页。

④ Rabassa, V., SIabri, O., Spaletta, C., “Conversational Commerce: Do Biased Choices Offered by Voice Assistants’ Technology Constrain its Appropriation?” October 22, 2021, https://doi.org/10.1016/j.techfore.2021.121292, September 10, 2024.

⑤ Courtl, R.,“Bias Detectives: the Researchers Striving to Make Algorithms Fair,” *Nature*, Vol. 558, No. 7710, 2018, pp. 357-360.

⑥ 陈昌凤、张梦：《由数据决定？AIGC 的价值观和伦理问题》，《新闻与写作》2023 年第 4 期，第 15-23 页。

同演化，彼此影响。生成式人工智能作为复杂社会技术系统的一部分，嵌入了社会结构中的权力关系、价值取向和文化偏见。其内生和外生偏见通过技术和社会的双向影响，不断互相强化和再生产。这一相互循环和强化过程说明，技术与社会并非简单的线性关系，而是通过多层次、多方向的反馈机制交织在一起。生成式人工智能的生成结果不仅会影响用户的认知和行为，还会反作用于社会环境，并且这种反作用通过反馈机制最终影响社会价值观的形成和变迁。随着这些变化逐渐积累并融入后续的数据和模型训练中，偏见向更广泛的社会层面扩散，形成偏见的累积效应。这种循环反馈机制使得生成式人工智能的偏见不再仅仅是技术问题，而是深深嵌入到社会的文化、政治和经济结构中，进一步加剧了现有的不平等和不公正现象。

三、生成式人工智能偏见的治理路径

生成式人工智能的崛起标志着人类科技发展进入到一个新的篇章，然而与此同时，正如技术之于文明有如利剑双刃，其所引发的社会风险与其带来的颠覆性变革如影随形。当未来生成式人工智能成为整个人类社会的“基础设施”，其内化的偏见和歧视将极大威胁全球社会的公平和公正。因此，探索生成式人工智能偏见的治理路径，对于保障这项新兴技术朝着健康和规范的方向发展，确保其设计和应用更好地符合社会的公平、公正和尊重价值，显得至关重要。生成式人工智能偏见的治理，需要构建“技术主体—用户主体—监管主体”多方协同治理的格局，以促进生成式人工智能发展与安全的平衡，实现人工智能与人类社会的和谐共生。

首先，从技术主体来看，需要在“对齐”（Alignment）原则的指导下，从技术流程与人员治理双管齐下，减少系统性偏见。所谓“对齐”原则，是指确保生成式人工智能的目标、价值和行为与人类价值观相一致和对齐的原则。[①] 在技术流程上，预处理数据、模型选择与后处理决策都应严格遵循该原则。具体而言，在数据预处理阶段，可通过采样、欠采样或合成数据等技术手段，确保数据能公平地代表包括被边缘化群体在内的广泛人口；在模型选择阶段，应采用基于群体或个体公平性的方法，如通过正则化或集成算法以减少偏见；在后处理决策中，通过调整模型的输出使其更加公正，如通过概率调整消除偏见。[②] 此外，开发人员的既有偏见也可能影响生成式人工智能的公平性，因此应该在开发和管理团队中强调技术的“道德编码”[③]，既要确保团队的多样性与包容性，又要强化技术伦理培训，让技术设计者和操作者明确其对潜在的偏见与歧视所应承担的责任。

其次，在用户主体层面，需要提升用户的媒介素养，并鼓励用户的积极参与。随着人工智能时代的到来，用户需掌握生成式人工智能的基本工作原理，并提升对其潜在社会风险的敏感度。这种素养不仅包括使用技术的能力，还应包含对技术背后隐含偏见的认知与抵御能力。[④] 通过信息反馈机制，用户可以主动参与纠偏。例如，通过用户与生成式人工智能的互动反馈，输入反偏见内容，优化模型性能，或通过建立反偏见数据集等方式，直接参与偏见治理实践。一个实际案

① Liu，Y.，Yao，Y.，Ton，J. F.，et al.，“Trustworthy LLMs：a Survey and Guideline for Evaluating Large Language Models'Alignment，” October 17，2023，http：//arxiv. org/abs/2308. 05374，September 10，2024.

② Ferrara，E.，“Fairness And Bias in Artificial Intelligence：A Brief Survey of Sources，Impacts，And Mitigation Strategies，” Vol. 6，No. 1，2023，p. 3.

③ 杨旦修、吕冠霖：《论智能传播机制及其价值风险治理》，《东南传播》2022 年第 6 期，第 36-39 页。

④ 厉晓婷、王传领：《人工智能时代用户媒介素养的养成：机遇、挑战及应对策略》，《中国编辑》2023 年第 10 期，第 74-78 页。

例是海地的民间数据组织与当地社区团体合作，建立了收集性别暴力数据的在线系统，从而形成了反偏见的数据集。①

最后，从监管主体来看，需要在保障和促进人工智能技术创新朝着更高更强的形态发展的前提下，建立动态校正的偏见监管体系②，推动技术的良性发展。2023 年 7 月 10 日，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局七部门发布了《生成式人工智能服务管理暂行办法》（以下简称《办法》），旨在促进生成式人工智能的健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益。在《办法》中，明确提出了“在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视”③ 的反偏见要求，体现了国家相关部门作为监管主体，高度重视生成式人工智能偏见带来的社会风险挑战，确保生成式人工智能技术朝着安全、公平、可控的方向发展。同时，世界其他不少国家和地区，以及欧盟等国际组织，也制定了相应的治理措施。例如欧盟在 2021 年提出《人工智能法案》，通过对人工智能的各类风险设定合规要求和提升透明度来应对其可能带来的偏见以及社会风险。未来，可考虑引入更为细化的偏见审查机制，尤其是在数据集和算法开发等关键技术阶段，强化审核和评估，并在相关法律法规的制定上引入公众参与机制，允许偏见受害者以及其他利益相关者就生成式人工智能的设计和实施发表意见，以提高人工智能立法的有效性和透明度。

生成式人工智能偏见的治理不仅是技术层面的挑战，更是对全球伦理和社会公平的严峻考验。有效的治理需要技术主体、用户主体和监管主体在“人类命运共同体”的愿景指引下，从全球视角展开多方协作和深入探索。全球合作的意义不仅仅体现在技术标准的协调上，更在于如何在尊重人类多样性与维护公平原则之间寻求动态平衡。这一治理过程要求人类社会重新审视技术进步与道德责任的内在关联，摒弃单纯追求经济增长或国家竞争的狭隘视角，转而迈向更加包容且富有责任感的全球治理模式，进而推动技术创新与伦理价值的深度融合，确保这一技术真正服务于全人类福祉。

（责任编辑：林春香）

① Smith，G.，Rustagi，I.，“When Good Algorithms Go Sexist：Why and How to Advance AI Gender Equity，” March 31，2021，https：//doi. org/10. 48558/A179-B138，September 10，2024.

② 刘艳红：《生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例》，《东方法学》2023 年第 4 期，第 29-43 页。

③ 国家网信办、国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局：《生成式人工智能服务管理暂行办法》，北京：中国法制出版社，2023 年。

Abstracts

The Action Logic，Risk Concerns and Comprehensive Governance of the Fan Support Phenomenon

YANG Jianyi，LIN Wenjun

Abstract：Fan support activities aim to rally behind a particular idol. From the phenomenon of fans attending TFBOYS' 10th anniversary concert in 2023，the "self-centered" support drive，the "platform-based circle group" support response，the "factional unity" support collaboration，the "spending competition" support rivalry，and the "capital-driven sponsorship" support dynamics constitute a complex action logic. As a result，the associated risks deserve attention：the ecosystem of public opinion communication becomes disrupted，the proper functioning of public order is affected，fan groups engage in conflict and hostility，undermining mainstream ideology，and endangering the physical and mental health of minors. To address the disorder in fan support activities，a comprehensive approach should be implemented. This includes building a harmonious spiritual environment through high-quality cultural offerings，expanding the dissemination of mainstream values with new technological tools，guiding fans to enhance their media literacy through multiple channels，normalizing disorderly support behaviors under the rule of law，and regulating capital's role with systematic supervision and norms.

Combining the Party's Self-Supervision and People's Supervision as a Potent Driving Force for Advancing the Party's Self-reform：Theoretical Foundations and Practical Innovations

CAI Zhiqiang，XU Li

Abstract：Leveraging the combination of self-supervision and people's supervision as a powerful driving force clarifies the dynamic mechanism，supervisory forms，and practical approaches of the Party's self-reform. This combination underscores the profound alignment between the Party's principles and the people's interests，refines the operational mechanisms of discipline inspection and supervision，and enriches the theoretical framework of discipline inspection and supervision. In the new era，in order to implement General Secretary Xi Jinping's important thought on the Party's self-reform and advance the integration of self-supervision and people's supervision，it is crucial to scientifically understand and fulfill the political requirements，institutional mechanisms，and methods for combining the two. It is essential to integrate intra-Party supervision with other forms of supervision，strengthen the institutional and regulatory system for the Party's self-reform，establish a supervision system with Chinese characteristics，foster a unified force for supervision，and pioneer a new dimension where the Party's self-reform guides the great social revolution.

Bias in Generative Artificial Intelligence：Manifestations，Mechanisms and Governance Paths

LUO Xi，CAI Wenyi

Abstract：As generative AI demonstrates its transformative potential across diverse industries and

fields, the issue of bias that comes with it has become increasingly severe. In the cultural domain, generative AI reproduces a Western-centric cultural universality, ignoring the uniqueness of different cultures; in politics, generative AI exhibits obvious partisan, ideological, and geopolitical biases, reinforcing the unequal global political and economic power structures; in gender, generative AI amplifies inherent societal gender biases through defamation, stereotypes, recognition biases, and underrepresentation. These biases arise from the mutual reinforcement, amplification, and reproduction of endogenous and exogenous biases in an ongoing cycle. The governance of bias in generative AI bias is not only a technical challenge but also a severe test of global ethics and social equity. It requires the construction of a multi-party collaborative framework of "technology entity, user entity, regulatory entity". The technology entity should adhere to the "alignment" principle to reduce systematic bias, the user entity should enhance their AI literacy and engage in bias governance practice, and the regulatory entity should establish a dynamic bias correction and supervision system, to promote a balance between the development and security of generative AI, achieving its harmonious coexistence with human society.

Gendered Body Techniques: Kindergarten Toilet Practices

FAN Xuan

Abstract: The concept of "body techniques" has significantly advanced Durkheim's theory and inspired the sociology of the body to capture embodied experiences in contemporary society from a technical perspective. This article discusses several underdeveloped aspects of body technique theory. Based on theoretical revisiting and an empirical study of children's toilet practices in kindergartens, three key findings emerge: Firstly, body techniques are not merely a "prereflective state"; their acquisition involves a reflective process. Individuals must undergo a reflective stage, overcome the guidance of social goals and knowledge embedded in these body techniques, and ultimately achieve a "post-reflective/de-reflective" state. The difficulties children face in acquiring toilet techniques, along with their frequent operational failures, show that there may be inherent conflicts among the goals of body techniques, and certain critical aspects of body techniques cannot be fully acquired through external demonstration, observation, or training. Instead, they require internal coordination to be achieved through continuous trial, error, and adjustment. Secondly, as a daily embodied form of collective representation, body techniques serve not only as a condensation of collective emotions but also as an intermediary for their continuation and transmission. These techniques create specific mental states that are crucial for individuals to internalize and understand the social goals underlying body techniques. In the case of children's toilet training, teachers' supervision and their expressions of negative emotions play a pivotal role. Thirdly, the gendered logic of body techniques in kindergartens shows that while there are biological differences between sexes, the way these differences are expressed in specific social contexts is shaped by gendered body techniques. As collective representations of gender norms, these techniques enforce the use of physical differences in ways that conform to established gender norms within a given social context.