认知不公视域下大语言模型幻觉的生成与治理*

高斯扬

（哈尔滨工业大学（深圳）马克思主义学院，广东 深圳 518055）

[摘 要]大语言模型作为对人类认知的技术性模拟，在复制人类知识生产机制的同时，也系统性地复制并放大了人类社会既有的认知不公。从认知不公理论视角审视，大语言模型幻觉的生成源于技术与社会两个相互强化的层面。技术层面，训练数据的代表性失衡与算法设计中的价值嵌入，导致模型在信息处理时出现结构性偏差，产生看似合理却偏离事实或公正性的输出；社会层面，技术开发与应用往往由少数企业或精英群体主导，其认知框架与利益偏好被编码进系统，从而固化了模型中已有的不公正知识的传播与再生产。二者交织作用，使大语言模型幻觉不再仅是技术缺陷，更成为一种深嵌于技术架构与社会权力结构中的认知压迫。治理大语言模型幻觉，需超越单纯的技术优化路径，转向构建一种融合认知正义的综合性治理框架。应在技术设计与训练中主动嵌入认知平等与公正原则，通过多样化数据采集、算法审计与价值对齐机制，从源头抑制偏见再生产；建立覆盖数据、算法、部署、反馈的全周期治理体系，推动跨学科协作与多元主体参与，确保技术发展符合公共利益；提升公众的数字素养与批判性认知能力，通过教育与社会对话削弱对模型的盲目依赖，重塑人在人机协同中的认知主体性。

[关键词]大语言模型 幻觉 认知不公 科技伦理 算法治理

[中图分类号]C02 [文献标识码]A [文章编号]2096-983X（2026）02-0130-09

一、引言

随着新一代人工智能技术的快速发展，大语言模型（Large Language Model）的“幻觉”（hallucination）成为一个不容忽视的问题。大语言模型幻觉是指模型在输出内容时出现事实错误、逻辑混乱的状况。尽管学者们采取多种技术手段来缓解大语言模型的幻觉问题，但幻觉难以被消除，甚至在某些应用场景中的出现频率较高。[1]事实上，大语言模型幻觉背后有着更为深层且复杂的原因，是“计算系统的固有局限”[2]。英国学者米兰达·弗里克（Miranda Fricker）①从人类社会固有的认知不公正角度，分析了人类知识在生产与传播过程中因身份偏见和解释资源分配不均导致的认知错误，提出了认知不公理论。[1]这一理论为探究大语言模

收稿日期：2025-09-08；**修回日期**：2026-01-08

*__基金项目__：国家社会科学基金青年项目“数字劳动助推新质生产力发展的理论逻辑与实践路径研究”（24CKS046）

作者简介：高斯扬，博士、副教授、博士研究生导师，哈尔滨工业大学（深圳）二十一世纪中国研究中心研究员，主要从事智能社会与数字劳动等研究。

①米兰达·弗里克在《认知不公：权力与认识伦理》中，首次提出了认知伦理理论。该理论从社会认识论与伦理学交叉的视角，系统分析了人类知识生产与传播过程中因身份偏见，如种族、性别、阶级等和集体解释资源分配不均而导致的系统性认知扭曲。参见：Fricker Miranda.Epistemic Injustice: Power and the Ethics of Knowing[M]. Oxford: Oxford University Press, 2007.

型幻觉背后更深层的认知伦理因素提供了深刻的启示。

二、认知不公：一个分析大语言模型幻觉的伦理视角

（一）对大语言模型幻觉进行认知不公分析的缘由

大语言模型（Large Language Model，LLM），是指在包含百亿及以上参数上预训练并能广泛应用于下游任务上的模型。[3]这种模型在语言能力、应用范围和智能程度等方面最具代表性，产生了如GPT-4、DeepSeek-R1等具有广泛影响力且被大规模应用的商业模型。大语言模型的幻觉指其生成的内容偏离了用户输入的任务和先前的上下文语境，输出了看似合理但与事实相悖的回答。[4]按照幻觉内容的性质，大语言模型幻觉被分为三种类型：一是语境幻觉，即模型生成的回答脱离了用户的意图或对话背景，与用户输入的任务不匹配。二是逻辑幻觉，即模型生成的回答中出现逻辑漏洞，比如推理链条断裂或因果倒置。[5]三是事实幻觉，即模型生成的回答中包含与已知事实不符的信息。

学界对大语言模型（LLM）幻觉问题的认知经历了一个从技术修复到伦理反思的深化过程。早期研究多聚焦于信息技术领域，将大语言模型幻觉视为技术缺陷，认为其根源在于模型性能不足，比如“训练数据的质量水平可能导致模型漂移（drift）或过度拟合（over-fitting）的问题，会显著减损大模型的预测能力”[6]。在此视角下，学者认为治理这一问题需要技术上的“幻觉探测—幻觉缓和”[7]，即针对模型的数据、算法和架构进行优化。在数据方面，学者们通过去除训练数据中的噪声、错误和偏见信息来减少大语言模型幻觉的来源。在算法方面，学者通过检索增强生成（Retrieval-Augmented Generation）和多模态验证[8]，利用多种模态的数据，如文本、图像、音频等，对算法的输出进行验证，进而提高输出内容的可信度。在模型架构方面，学者们利用解码策略优化（Decoding Strategy Optimization）策略，调整模型在生成文本时的搜索策略和选择机制，降低生成错误内容的概率。以上方法在一定程度上有效，但无法消除大语言模型幻觉。这使学者们产生了“大语言模型幻觉是无法消除的”[2]的判断。

近两年，伦理学者将幻觉问题置于更广阔的社会认知结构中加以审视，提出了认知伦理路径。伦理学者指出，大语言模型幻觉实质上是人类社会不平等的认知权力关系在技术中的投射。Safiya Umoja Noble[9](P3)、刘永谋[10]等通过研究发现，人类社会的种族、性别、阶级等不平等权力会深刻影响人类认知，大语言模型会继承并放大这些社会认知权力的不平等。大语言模型幻觉不应仅被当作技术故障，而应被视为一种认知不公视角下的认知偏差。

（二）认知不公视域下大语言模型幻觉的内涵

认知不公是探讨知识生产与传播过程中因社会身份偏见导致社会认知偏差的理论。这一理论揭示了社会权力不平等如何系统地扭曲我们的认知实践，降低认知的准确性与真实性，为理解大语言模型幻觉提供了关键的分析框架。事实上，人类社会中不平等的权力关系会导致特定个体或群体的知识贡献被不合理地贬低甚至忽视。这种忽视会使得这些个体或群体在知识的生产、传播与接受过程中处于不利地位，形成一种系统性的认知障碍。弗里克将这种认知障碍的来源分为证言不公和解释不公两个方面。

证言不公指的是由于听者对说话者的社会身份偏见，导致说话者的证言可信度被不公正地降低，其知识主张难以得到应有的认可和采信。比如在小说《杀死一只知更鸟》中，黑人汤姆·罗宾逊因种族偏见而被陪审团无视其真实证词，最终被错误定罪。解释不公是指边缘群体因集体解释资源的缺乏，无法使其重要的社会经验被充分理解。而社会中的解释工具，如概念、词汇、叙事框架往往被“社会制度和实践中有权有势的人”[11](P148)掌控，这导致了“社会理解方面的不公平”[11](P148)。在缺乏足够的概念

工具和话语体系来理解、表达和传达自身的经验和困境的情况下，一些社会群体的独特经历难以被社会所理解和承认，无法有效地参与到社会认知对话中，这种认知偏差源于社会认知层面的结构性不公。

将弗里克的认知不公理论引入对大语言模型幻觉的分析，我们可以更清晰地洞察到，大语言模型幻觉不仅仅是模型在技术层面上对事实的偶然偏离或逻辑的随机混乱，而是植根于社会认知权力结构的不平等。大语言模型作为对人类语言统计规律的技术化模拟[12]，其在设计、运行和应用过程中会将人类社会存在的认知不公现象复制并表现出来。一方面，大语言模型是通过海量数据训练、对人类语言统计规律进行技术化模拟的系统，其训练数据本身就深刻嵌入了人类社会主流群体的视角和解释框架。如果边缘群体的独特经验、地方性的知识未在数据中得到充分表征，这些知识便很难在大语言模型的输出中被准确理解和表达。另一方面，模型的算法设计、优化目标以及部署策略，会受到开发者、商业企业等特定主体认知视角的影响，这些主体的社会权力和认知偏好可能在模型中固化，进一步放大既有的认知不公。

（三）大语言模型幻觉对认知不公理论的扩展

认知不公理论揭示了人际层面基于身份权力的认知压制关系，但大语言模型幻觉的大规模出现，表示着这种人际层面的认知不公正扩展到了技术应用的层面。事实上，在大语言模型的作用下，认知关系中的强者并非直接作用于弱者，而是通过大语言模型这一技术载体来实现其影响力。

具体而言，大语言模型凭借其强大的计算能力和广泛的数据处理，在知识生成和信息传播方面具有巨大作用。当掌握技术与数据的人或机构通过对模型的训练数据选择、算法设计和参数调整等，将自身的认知偏好和利益诉求嵌入到大语言模型中时，普通用户在使用大语言模型获取信息时，接收到的内容往往是经过预设的、带有特定倾向的，这使得普通用户在认知过程中受到无形的限制和引导。换言之，大语言模型在设计阶段与认知关系中的强者结合而携带强者的认知偏见，使其携带了认知不公的基因。而在应用阶段，当大语言模型与强势方的利益和视角耦合后，它不再是中立的工具，而将成为一个能够高效、大规模地再生产与强化这些不平等的认知技术。这不仅仅是对弗里克认知不公理论的简单复制，而是一种技术性的扩展重构。即当强者利用大语言模型获得了前所未有的认知支配规模与效率，使得弱者在更隐蔽、更难以挑战的系统性偏见面前，陷入了更深层的认知困境当中。换言之，大语言模型作为“与人类实践相结合的社会技术实体”[13]，其“每一种技术架构、每一行代码、每一个界面，都代表着选择，都意味着判断，都承载着价值”[14]。传统的、人际间的认知不公，演变为一种由技术架构和算法逻辑所承载的隐性权力作用关系。大语言模型会内化和再现人类社会固有的证言不公与解释不公，与社会权力中的强势方相结合，导致认知权力的集中和弱势方的认知依赖，削弱了个体作为认知主体的地位。

三、认知不公视域下大语言模型幻觉的生成

从认知不公的视角出发，大语言模型幻觉是技术层面的数据不公、算法不公与社会权力结构相互作用的结果。其中数据不公与算法不公导致模型在信息处理内嵌了系统性偏差，而开发者、企业主导了大语言模型技术的部署与应用，固化了已有的社会认知偏见。以上因素相互作用，使大语言模型幻觉不断出现。

（一）数据不公和算法不公

数据不公与算法不公是导致大语言模型产生认知不公的两个核心技术根源，二者共同强化了大语言模型的系统性偏见。数据不公是指在大语言模型的训练过程中，数据的选取、标注和使用存在偏差。事实上，训练数据的质量

直接决定了大语言模型的输出效果，然而训练数据的收集、处理和标注等环节，均会出现与证言不公相似的数据的不公正。[7]以GPT-3和LLaMA2为例，GPT-3的训练数据集中“英语语料占据了92.65%，欧洲各国语言的总占比超过5%，汉语语料的占比不到0.1%”；在以多元化为目标的LLaMA2训练数据中“汉语语料占比为0.13%”，非洲语言、原住民语言等几乎处于数据盲区。[15]这种数据代表性的严重失衡，意味着人类知识体系在数据层面就受到了不平等的对待，直接导致模型在认知模仿过程中，弱势群体的经验被边缘化，其声音在生成的知识中近乎湮没。

算法不公是指大语言模型的算法设计和运行过程中，因数据偏见、技术机制等因素，导致其输出结果系统性损害特定群体公平性的情况。算法作为大语言模型的核心驱动力，其公正性直接关系到模型输出的准确性和客观性。然而，算法需要数据来训练，数据的偏见会导致模型在识别和理解语言时，出现类似解释不公的认知偏差。例如，当训练数据中缺乏或者排除了边缘群体的真实经验，由数据训练出来的算法就会将这种被排除后的经验编码为正常、标准的认知框架。以肤色算法为例，2018年麻省理工学院媒体实验室研究员Joy Buolamwini与微软科学家 Timnit Gebru指出，三款主流算法对深肤色女性的分类误差率最高可达34.7%，远高于浅肤色男性的0.8%。[16]原因在于，算法训练数据中深肤色女性的样本严重不足。这会让算法高度依赖高频出现的浅肤色男性的数据特征，将浅肤色男性的面部参数默认为基准值，造成算法解释这些数据时出现潜在认知偏见。此外，大语言模型基于统计规律的处理机制，常将高度复杂、多维的人类文化、价值观等问题进行简化计算，这种技术逻辑上的“降维”处理会引发认知错误。尽管计算理论（如丘奇-图灵论题）指出了计算的边界[17]，但算法在实践中仍倾向于处理可计算的对象，而将不可完全计算的社会复杂性强行简化，这本身就是一种认知强权的表现。

数据不公与算法不公并非孤立存在，二者构成一个相互加强的循环。一方面，有偏见的数据训练出有偏见的算法，算法又将这种偏见编码为看似客观的认知框架。另一方面，有偏见的算法在处理现实世界时，会主动发现并收集更多能验证其偏见的数据。以美国的警务预测算法为例，这一算法包含了大量反映历史种族偏见的因素[18]，这导致其对某一黑人社区的犯罪风险评分过高。而过高的犯罪评分使警务算法在预测该社区的犯罪率时，会进一步探测到更多的疑似犯罪的数据，这使原本带有偏见的算法得到了数据方面的验证，形成了“数据偏差-算法偏差”的闭环强化。这一循环极大地加剧了大语言模型的幻觉现象，使其输出的不公内容显得愈发“真实”和“客观”。

（二）开发者、商业企业和普通用户的不平等关系

在大语言模型的应用场景里，开发者、商业企业和普通用户之间存在着显著的不平等关系，这会加剧大语言模型的幻觉情况。首先，大语言模型的开发者处于技术核心地位，他们掌握着大语言模型的研发技术、算法设计和数据选择等关键环节。开发者会根据自身的认知和偏好来决定模型的功能、特性以及适用范围。例如前文提到Chat GPT-3和LLaMA2。这两个模型训练数据过度偏重英文文献与西方文化场景，导致二者排斥了非西方社会的认知经验和多元数据，降低了其关于非西方社会知识的准确性。实际上，这些模型的开发者将西方学术话语体系与主流数据标准默认为普适真理，这种认知偏好塑造了上述被大范围、大规模使用的大语言模型的认知倾向，使得这些模型在处理非西方文化或边缘群体相关问题时极易出现认知偏差。

进一步地，商业企业是大语言模型的推广者和盈利者，他们会因为在资源和市场等方面的实际需要，降低大语言模型的认知公平性。事实上，大语言模型从技术发明走向产业化和

商业化，需要大量的预付资本才能实现。而风险资本是推动技术发展的主要资金来源。一方面，风险资本为技术企业提供了发展大语言模型的大量资金，推动了技术企业在数据处理技术、算法技术等方面的创新和应用。另一方面，风险资本通过投资技术企业来寻求高额的回报。这种高额回报并不来自这些企业的当下利润，而是依赖于它们在未来的盈利预期，这使“大模型的技术价值取向在产业化阶段会受到平台管控者、商业资本的影响”[7]。以Grok 3模型为例，这一模型在处理政治争议话题时，因投资方与特定利益集团相关，会对敏感信息采取回避事实的差异化回应。①Open AI用GPT-3向世界展示了千亿级参数的技术优势，却未充分关注并解决其内在的数据偏见问题，导致模型在生成内容时大量输出幻觉。美国学者Bender 等人批评这种只重视参数“越大越好”但不重视参数治理的企业偏好，指出其本质是技术企业通过参数壁垒来巩固市场地位。[19]商业企业受到逐利需求的推动会在一定程度上将自身利益偏好置于模型的认知公平之上，损害大语言模型的认知公正性。

与之相区别的是，普通用户在使用大语言模型的过程中处于被动地位。普通用户作为大语言模型的终端使用者，本应与大语言模型相互促进、共同发展。然而，在当前技术框架下，普通用户被排除在了大语言模型的认知过程之外。技术企业通过技术壁垒与制度性约束两种手段，排除普通用户在大语言模型应用过程中的认知参与。技术壁垒来自大语言模型的复杂技术架构和不透明的算法逻辑。比如，因ChatGPT系列大语言模型采用了闭源架构，普通用户不仅无法查看其模型的源代码，而且无法知晓模型内部的决策逻辑和数据处理过程。同时，由于拥有这些模型的商业企业不公开训练数据的来源、清洗过程和筛选标准，使用户更加无法知晓数据中是否包含偏见、是否充分覆盖边缘群体的认知需求，这种不透明性切断了用户参与数据筛选和模型优化的可能性，导致用户难以对大语言模型输出的公正性进行有效监督。制度性约束是指技术企业会使用不平等的服务条款来限制用户参与知识决策的权利。在企业特定服务条款的规定下，用户仅能以数据提供者的身份使用模型，而无法对模型设计、训练逻辑提出异议或参与改进。这种排斥使用户在使用过程中处于“数字佃农”[20]的处境。这影响了普通用户对大语言模型应用的认知参与，降低了大语言模型在应用中的公正性，加剧了大语言模型的认知不平等。

（三）技术层面认知不公和社会层面认知不公的相互强化

大语言模型在技术层面与社会层面的认知不公并非孤立存在，而是相互交织、彼此强化的关系。其中，大语言模型在技术层面的认知不公是社会不公在数据、算法等领域的投射与固化，而社会权力结构则决定了技术不公的生产逻辑与修正边界，二者共同构成一个难以打破的闭环。

从技术层面来看，训练数据的选择偏差和算法设计中的隐性偏见共同导致模型在知识获取阶段就存在系统性偏差，这些偏差往往来自现实社会中既有的不平等现象。以性别偏见为例，大语言模型并不真正理解什么是性别偏见，只是按照数据里的高频出现模式，机械地把某个性别和某种偏见标签关联在一起。开发者发现这种错误关联后，通常会通过调整参数和算法的方式来避免大语言模型产生明显的偏见输出。但这种机械主义的调整方式，只会使大语言模型更加偏离数据偏见背后的深层次社会因素，加剧大语言模型的“随机鹦鹉”现象[1]，强化其内在的社会偏见。

从应用层面来看，大语言模型的开发者、商

①2025年2月，有用户报告称，当被问及“谁是最大的虚假信息传播者?”并启用“思考”设置时，Grok 3 在其“思考链”中指出，它被明确指示不要提及埃隆·马斯克。马斯克是Grok 3的投资人。参见：胡泳，王昱昊.技术过程论视角下AI幻觉生成的价值负荷与伦理问题探析[J]. 南京社会科学，2025（3）：84-94.

业企业与普通用户之间存在显著的权力与知识不对称。强势方（如科技公司）不仅能利用技术优势定制符合自身利益的模型输出，还拥有定义何为“偏差”及如何“修正”的话语权。相反，普通用户因缺乏技术和资源，不仅难以识别和挑战模型的系统性偏差，反而可能将大语言模型充满偏见的输出视为“提供各种解决方案、中立和公正”[12]的认知代理。这种认知的代理和依赖关系会使用户的批判性思维逐渐钝化，进而陷入认知被动状态。

最终，社会不公塑造了大语言模型中有偏见的数据，从而生产出带有偏见的技术系统；而这套技术系统在应用中又被社会权力结构中的强势方所掌控，其输出结果反过来“验证”和“合理化”原有的社会偏见。在这一循环中，大语言模型从一个技术中立的工具，演化为一个主动生产并固化偏见的“引擎”，使认知不公从人际间的偶然现象上升为一种嵌入技术架构的、系统性的认知压迫。

四、认知不公视域下大语言模型幻觉的治理

习近平总书记指出，“要把握人工智能发展趋势和规律，加紧制定完善相关法律法规、政策制度、应用规范、伦理准则，构建技术监测、风险预警、应急响应体系，确保人工智能安全、可靠、可控”。[21]面对大语言模型幻觉所反映的认知不公问题，需明确大语言模型发展的价值导向，将认知平等和正义的价值导向融入大语言模型的全生命周期体系，并提升公众的大语言模型素养。

（一）明确大语言模型发展的价值导向

坚持技术向善，是治理大语言模型幻觉中认知不公的关键所在。平等和公正是我国社会主义核心价值观的重要内容，这些内容落实在我国大语言模型领域就是要确立大语言模型在认知领域的平等和公正。为了实现这一目标，要建立一套清晰的大语言模型发展的价值导向框架，明确大语言模型发展的伦理边界。例如，在数据采集阶段，应主动纳入更多元化的语料来源，避免单一文化或群体的过度代表；在算法设计中，需引入公平性评估机制，确保模型输出不会系统性地损害特定群体的利益。此外，还应通过跨学科合作，吸纳社会学、伦理学、法学等领域的专家参与技术决策，以弥补技术开发者可能存在的认知盲区。

同时，明确大语言模型发展的认知伦理还需要从制度层面入手，推动行业规范和法律法规的完善。尽管我国针对包括大语言模型在内的生成式人工智能技术，先后发布了《生成式人工智能服务管理暂行办法》和《生成式人工智能服务安全基本要求》。但这些办法和要求仅对技术企业“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”[22]的部分负面行为（如企业侵犯用户的数据隐私、算法歧视等）做了规定，而没有对技术企业应承担的正向认知责任进行系统阐述。确保大语言模型的技术向善发展，应对大语言模型在促进认知平等和公正的积极作用进行顶层设计和整体规划。尤其要明确技术企业在算法开发、数据训练、模型应用等环节中应承担的认知平等和公正的责任，确保大语言模型发挥积极的认知平等和公正促进作用。

还要完善大语言模型发展的伦理规范、行业自律。行业协会、企业联盟是由大语言模型开发市场主体构成的，他们是确保大语言模型促进认知平等和公正的“第一责任人”。通过行业公约、企业规范，可以让技术企业和智能产业在追求经济效益的同时，积极履行社会责任，将促进认知平等和公正纳入大语言模型的发展议程，共同维护良好的技术生态，发挥大语言模型促进认知发展的正向作用。

（二）将认知平等和正义的价值导向融入大语言模型的全生命周期体系

大语言模型中促进认知平等和公正的底层能力来自其数据训练和算法设计。“从根本上说，提高训练数据的质量和多样性是改善大

模型幻觉、完善人机传播流畅性的基础。”[23]治理大语言模型幻觉中的认知不公问题，要抓住大语言模型开发的数据训练和算法调整两个关键环节，从数据过滤与算法审查入手，强化技术监管。

鉴于西方开发企业在GPT、LLaMA、Grok等大语言模型训练中对英语语料的偏重而导致的认知不公情况，有学者提出，应建立具有中国文化主体性的数据库[24]。事实上，中国拥有全球规模最大的数据资源，具备为建立反映我国认知平等和公正价值导向的数据库，提供丰富且安全的数据资源的条件。可以由政府牵头，以国企为依托，建立具有中国特色的高质量数据库。其中，建库企业可在数据收集和整理的阶段采用半自动的人工标注或者自动标注，将包含认知平等和公正在内的社会主义核心价值观编码为生成式人工智能的价值基因。但需注意的是，建立这种数据库要严格遵守《中华人民共和国个人信息保护法》《中华人民共和国网络安全法》等法律法规在个人信息保护方面的要求，遵守《生成式人工智能服务安全基本要求》《生成式人工智能服务管理暂行办法》等对语料来源、内容以及标注等环节的相关规定。

要推动大语言模型算法的完善与优化。具体而言，大语言模型的开发企业与技术人员需秉持算法设计的价值敏感性，把包含平等和公正在内的认知伦理体系转化为具体的伦理规范和技术要求，并主动将其融入算法目标设计、数据选择、结果呈现的全流程，形成“自上而下”的价值对齐技术系统；同时要强化大语言模型内部算法审计，通过“自下而上”的方式，对大语言模型在设计、开发、生成、应用等环节的算法决策过程“进行溯源审计与治理”[25]。此外，政府部门要建立算法审查机制，以推动认知的平等和公正为标准，通过关键词、分类模型、人工抽检的方式，消除企业模型中传播虚假信息、暗含歧视观点的有害内容。

要建立和完善认知伦理治理相关的法律法规体系。法律法规要包含大语言模型的设计者、使用者和维护者等各种治理主体的法定权责，划清各治理主体的责任边界与权力清单，对认知平等和公正问题给予具体清晰的界定、阐释与规约。这样才能为大语言模型领域提供更多的行业依据与管理保障。同时，政府和相关管理机构应鼓励技术开发者和技术企业采用更加开放和透明的技术架构，“在技术发展伊始就将与生成式AI技术发展有关的各个社会群体都纳入技术的设计过程”[26]，促进开发者和技术企业吸纳不同社会群体的认知经验；还要鼓励技术开发者和技术企业积极探索新的算法设计，使大语言模型在追求效率和准确性的同时，更多实现认知的平等和公正。

（三）提升公众的大语言模型素养

大语言模型的素养是用户运用大语言模型进行问题分析、知识检索和答案生成的能力，这种能力越高，用户愈能提升对大语言模型的使用能力，抵抗大语言模型应用过程中不平等关系的作用，减少大语言模型幻觉对自身的负面影响。

要让公众掌握大语言模型的训练方式和输出机制。只有让公众明确了解模型的工作原理和可能存在的偏见，才能分析和判断幻觉。这就需要政府调动教育部门、社会公益组织和高校等社会资源。比如，教育部门可以设置一定的实践活动，让学生了解大语言模型的工作原理、潜在风险及幻觉类型，从而培养他们的辨识能力和批判性思维；社会公益组织可以向公众介绍大语言模型的研发背景，讲解不同企业大语言模型的运行方式和过程，让公众以谨慎的态度使用大语言模型；高校可以借助自身的科研力量，推出一系列计算机和网络行业的专家关于大语言模型引发风险的线上主题课程和讲座，增强公众对大语言模型幻觉、信息茧房和认知不公等方面的认知。

同时，政府部门也要积极转换大语言模型在设计和应用过程中产生认知不公问题的应对思路。大语言模型在人工智能时代给认知活动带来的既是挑战，也是机遇。抓住机遇，转变思

路积极应对，才能使大语言模型更好地服务于人类认知活动的发展。政府可以通过鼓励政府部门、学术界和产业界合作，共同探索如何利用大语言模型提升认知的平等和公正；可以通过举办论坛、研讨会等活动，促进各方经验分享和智慧碰撞，凝聚行动合力。此外，政府和相关管理部门要积极探索人与技术协同进化的新模式，构建开放、包容、公正的技术生态。

五、结语

大语言模型作为强大的认知技术中介，重构了人类的认知权力。传统的、基于身份偏见的证言不公与解释不公，被内化并系统性地编码进模型的训练数据和算法逻辑之中。这使得认知不公从一种人际间的偶然现象，升级为一种嵌入技术基底、可大规模复制的系统性偏差。社会既有的不平等在数据采集和算法设计阶段被植入大语言模型，在应用过程中又因其输出被社会权力结构中的强势方，如技术精英、商业资本支配，从而进一步放大和固化这些不平等。技术层面的认知不公与社会层面的权力不对称相互加强，共同构成了一个难以打破的闭环，使得大语言模型幻觉中的认知不公成为一个根植于技术、扩大于社会、作用于个体认知的复杂系统性问题。大语言模型环节警示我们，不能将人工智能的伦理问题简单视为技术缺陷，而必须从技术哲学、社会伦理与治理政策的交叉视角进行综合治理。

大语言模型的幻觉治理应超越单纯的技术修补，迈向更深层次、更跨学科的探索。未来研究要关注能够揭示模型决策逻辑、特别是其价值判断与偏见来源的可解释人工智能技术，为公平、正义等伦理概念设计可计算、可审计的度量指标。还要将伦理价值前瞻性地嵌入技术设计的全生命周期。更要探索建立有效的公众反馈与算法审计渠道，“构建更加公平、包容和可持续的技术发展路径”[27]。

应对大模型时代的认知不公并非要追求一个全然无偏的完美的人工智能技术，而是要开辟出一条负责任的、可抗辩的、促进而非削弱人类认知主体性的人机协同新道路，这需要技术开发者、政策制定者与全体社会成员共同承担起责任。

参考文献：

[1]刘永谋. 警惕AI“幻觉”带来的安全风险[J]. 科学大观园，2025(7)：58-61.
[2]XU Z, JAIN S, KANKANHALLI M. Hallucination is inevitable: Aninnate limitation of large language models[J/OL]. (2024-01-22)[2025-11-22]. https://doi.org/10.48550/arXiv.2401.11817.
[3]矣晓沅，谢幸. 大模型道德价值观对齐问题剖析[J]. 计算机研究与发展，2023(9)：1926-1945.
[4]赵月，何锦雯，朱申辰，李聪仪，张英杰，陈恺. 大语言模型安全现状与挑战[J]. 计算机科学，2024(1)：68-71.
[5]GUAN X, LIU Y, LIN H, LU Y, HE B, HAN X, SUN L. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting[J/OL]. (2023-11-22)[2025-11-22]. https://doi.org/10.48550/arXiv.2311.13314.
[6]曾雄. 人工智能大模型价值对齐的现状考察、问题检视与规范进路[J]. 电子政务，2025(2)：34-44.
[7]胡泳，王昱昊. 技术过程论视角下AI幻觉生成的价值负荷与伦理问题探析[J]. 南京社会科学，2025(3)：84-94.
[8]储节旺，周柯堰. 知识增强生成赋能知识生产模式变革的研究[J]. 现代情报，2026(1)：14-24.
[9]NOBLE S U. Algorithms of Oppression[M]. New York: New York University Press, 2018.
[10]刘永谋，彭家锋. 算法决策的认识论不公正及其矫正[J]. 山西大学学报(哲学社会科学版)，2023(6)：1-9.
[11]FRICKER M. Epistemic injustice: Power and the ethics of knowing[M]. New York: Oxford University Press, 2007.
[12]BROWN T B, MANN B, RYDER N, etal. Language models are few-shot learners[EB/OL]. (2020-05-28) [2025-11-22]. https://doi.org/10.48550/arXiv.2005.14165.
[13]GAl M S. Algorithms as illegal agreements[J]. Berkeley Technology Law Journal，2019(1)：67-118.
[14]用主流价值纾解“算法焦虑”[N]. 人民日报，2018-06-20(5).
[15]苗逢春. 生成式人工智能及其教育应用的基本争

议和对策[J]. 开放教育研究, 2024(1). 4-15.
[16]BUOLAMWINI J, GEBRU T. Gender shades: Inter sectional accuracy disparities in commercial genderclassification[J]. In Conference on Fairness, Accountability and Transparency. 2018(81): 77-91.
[17]刘晓力. 计算主义质疑[J]. 哲学研究, 2003(4): 88-94.
[18]高宇航, 白惠仁. 数智社会中的结构性认知非正义[J]. 科学学研究, 2025(8): 1623-1631.
[19]BENDER E M, GEBRU T, MCMILLAN-MAJOR A, SHMITCHELL S. On the dangers of stochastic parrots: Can language models be too big?[J]. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021(5): 610-623.
[20]巩永丹, 李彦甫. 走出数字资本主义“封建化”的迷思——“技术封建主义”思潮的逻辑要旨及其理论检审[J]. 马克思主义理论教学与研究, 2025(1): 64-75.
[21]共产党员网. 习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展[EB/OL].(2025-04-26)[2025-11-22]. https://www.12371.cn/2025/04/26/ARTI1745635107727513.shtml.
[22]生成式人工智能服务管理暂行办法[EB/OL].(2023-07-10)[2025-11-22]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
[23]张铮, 刘晨旭. 大模型幻觉: 人机传播中的认知风险与共治可能[J]. 苏州大学学报(社会科学版), 2024(5): 171-180.
[24]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[J]. 江苏社会科学, 2024(1): 66-74, 242.
[25]袁雨晴, 陈昌凤. 道德物化: 大模型人机价值对齐的技术伦理进路[J]. 南京社会科学, 2024(6): 88-97.
[26]胡明艳, 王少华. 生成式 AI 的认知不公问题及其纾解[J]. 科学学研究, 2025(10): 2232-2240.
[27]徐涛. 数字社会主义理论批判性分析[J]. 深圳社会科学, 2026(1): 121-129.

【责任编辑 邱佛梅】

The Generation and Governance of Large Language Model Hallucinations from the Perspective of Epistemic Injustice

GAO Siyang

Abstract: As a technical simulation of human cognition, large language models not only replicate the mechanism of human knowledge production but also systematically replicate and amplify the existing epistemic injustice in human society. From the perspective of the theory of epistemic injustice, the generation of large language model hallucinations stems from two mutually reinforcing levels: technology and society. At the technical level, the imbalance in the representativeness of training data and the value embedding in algorithm design lead to structural biases in the model's information processing, resulting in outputs that seem reasonable but deviate from facts or fairness. At the social level, the development and application of technology are often dominated by a small number of enterprises or elite groups, and their cognitive frameworks and interest preferences are encoded into the system, thus solidifying the spread and reproduction of existing unjust knowledge in the model. The interaction of these two factors makes large language model hallucinations not just a technical defect but a cognitive oppression phenomenon deeply embedded in the technical architecture and social power structure. To address large language model hallucinations, we need to go beyond the simple path of technical optimization and shift towards constructing a comprehensive governance framework integrating epistemic justice. The principles of epistemic equality and fairness should be actively embedded in technical design and training. Through diverse data collection, algorithm auditing, and value alignment mechanisms, we can suppress the reproduction of biases at the source. A full-cycle governance system covering data, algorithms, deployment, and feedback should be established to promote interdisciplinary collaboration and the participation of multiple stakeholders, ensuring that technological development serves the public interest. We must improve the public's digital literacy and critical cognitive ability. Through education and social dialogue, we can weaken the blind reliance on the model and reshape human epistemic subjectivity in human-machine collaboration.

Keywords: large language model; hallucination; epistemic injustice; technology ethics; algorithm governance