“AI 幻觉”的社会影响与多维治理

吴源泰[1]， 袁世明[2]， 刘欣怡[2]

（1 大连东软信息学院，沈阳 大连 2 四川科宏石油天然气工程有限公司，四川 成都 ）

摘要：

幻觉作为人工智能技术的一种特性，是 AI 识别、生成等各种超能力的的根源之一，随着生成式 AI 的发展，幻觉也作为一种技术缺陷，成为影响着社会发展变量。本文重点剖析了“AI 幻觉”的生成机理，探讨其对信息生态、社会运行、技术发展及人类进程的多维影响，最终从技术优化、制度规制、应用层和社会层四个层面构建协同治理系统，为推动 AI 技术安全可控发展提供参考。

关键词：AI 幻觉；大模型技术；生成式 AI；社会风险；协同治理；多维治理

1. 绪论

1995 年，美国计算机科学家史蒂芬·泰勒（Stephen Thaler）首次引入了“幻觉”这一概念，用于指代人工智能自发产生的新想法，这时的幻觉有一定的褒义色彩，幻觉意味着 AI 具备一定的创造力。2017 年，这个色彩发生了变化，幻觉≠创造力，幻觉被研究人员用来描述在使用翻译模型时出现错误的现象。2022 年 8 月，Facebook 的母公司 Meta 在论文中形容幻觉就是自信的说假话，这就是关于“AI 幻觉”的最新定义，可称之为“自信的谎言家”(徐剑,2025)。

三个月后，随着 ChatGPT 惊艳亮相，AI 从工具性应用迈向价值性渗透，生成式 AI 的崛起，极大的影响着人们的工作与生活，然而 AI 内在的“幻觉”问题，即生成虚假但看似合理的内容(李凌,2025)，使 AI 成为一把双刃剑行走在公共社会中。从德勤会计事务所使用 OpenAI GPT-4o 驱动的生成式 AI 生成虚假报告被重罚(Georgia Institute of Technology,2025)，到 AI 输出深度嵌入医疗诊断、司法辅助、新闻传播和教育等核心社会职能，“AI 幻觉”由一种技术奇观转化为一种社会风险。理解幻觉全方位影响并构建有效应对策略，已成为关乎社会稳定与可持续发展的重要议题。

本文通过解析幻觉形成的路径：构建数据库-训练大模型-推理解码，搭建“技术-制度-应用-社会”多维治理体系，搭建在发挥 AI 效能的同时，防控幻觉的社会风险系统。

2. “AI 幻觉”生成机理

“AI 幻觉”是大型语言模型等生成式 AI 在输出过程中，产生与可验证事实不一致、内部逻辑断裂或偏离输入上下文的内容现象。区别于主观恶意的误导行为，本质上是模型认知机制缺陷导致的无意识输出偏差。幻觉的产生并非偶然，而是数据库架构局限、训练数据局限与人机互动

作者简介： 吴源泰[1]，大连东软信息学院工学学士，研究方向：软件工程，邮箱：2729302286@qq.com
袁世明[2]，西南石油大学硕士研究生，研究方向：BIM、结构设计、油气地面工程建设，邮箱：1310440398@qq.com
刘欣怡[2]，西华大学工学学士，研究方向：给水排水工程，邮箱：516942626@qq.com

推理解码的必然结果。数据是幻觉的根源，知识源头呈现“慢性中毒”现象(徐延民,2025)。低质量的数据可能会引入偏见或错误，不全面的数据会导致 AI 的凭空捏造。数据正确训练环节的缺陷也可能导致 AI 运力不足，因架构缺陷，AI 可能无法捕捉复杂的上下文关系，而训练策略，可能导致 AI 不再参考标准答案，自行生成内容，在训练环节，还可能出现 AI 为了迎合人类偏好而营造幻觉的情况成为“Impostor Bias(冒名顶替者 Bias)”(Casu M,2025)。为了让 AI 输出内容不要千篇一律，解码就要有随机性，因此 AI 会主动使用一些冷门甚至不准确词汇，同时，有时 AI 又会过度追求文本的流畅性，导致准确性被忽略，最终幻觉成了这一轮大语言模型的通病。

3. “AI 幻觉”对社会发展的多维影响

3.1. 信息生态层面：知识体系污染

“AI 幻觉”通过“生成-传播-再训练”的循环污染信息生态。语言大模型自主衍生编造的虚假事实一旦进入传播渠道，不仅会被公众当作可靠信息接收，还可能被其他 AI 系统抓取纳入训练数据，大规模、低成本地生成高度逼真的虚假新闻、深度伪造内容和不实信息，形成“幻觉循环”，加剧了信息污染。从而陷入事事怀疑，凡是求证的地步。

3.2. 社会运行层面：冲击公共信任

在公共服务领域，幻觉已对治理效能产生实质干扰。政务部门虽明确 AI 的“辅助型”定位，但模型仍可能在政策解读中产生幻觉，导致公众误解政策内容，削弱政府公信力。2024 年，一名乘客起诉加拿大航空，起因是乘客通过该航空公司的聊天机器人了解的折扣信息，是机器人在训练过程中从其他航司学习得来的，而该航司并不存在这项折扣。这使公众陷入“真相困境”，难以辨别真伪，从而侵蚀社会共享事实的基础，进而社会凝聚力与民主决策过程也将遭到破坏。

3.3. 技术发展层面：制约 AI 应用与创新进程

中国信通院的调研显示，42.2%的公众反馈 AI 生成内容不准确，这种信任赤字严重阻碍技术落地。在金融、法律、工程设计等对精确性要求高的领域，企业因担忧幻觉风险不敢深度应用 AI，制约着数字化转型进程，进而影响创新方向的发展，使 AI 技术偏离“智能向善”的轨道，引发更广泛的是非争议。

3.4. 人类发展层面：人类认知与技能的退化

在工作和生活上，过度依赖 AI 作为“外部大脑”，将导致人类批判性思维能力和事实核查习惯的退化。当一代人习惯于接受 AI 提供的、未经甄别的答案，其独立探索知识、深度思考和严谨求证的核心能力可能被削弱，从长远看，这将对人类文明的创新潜力和智力发展构成深层挑战。

4. “AI 幻觉”的多维治理体系

4.1. 技术优化：从“数据校准”到“可靠性提升”

技术优化聚焦模型数据校准与外部验证强化。在模型研发端，优化训练数据，进一步调整训练方式，让幻觉出现的几率从源头上降低，采用高质量标注数据与合成数据相结合的方式，减少训练数据中的错误隐患。同时重构评估体系，引导模型形成“知之为知之，不知为不知”的行为模式，不随意臆想发挥污染数据。

可靠性提升可采用检索增强生成（RAG）技术，将模型与权威知识库实时对接，强制模型在事实性问答中进行来源核验。建立多模型交叉验证机制，对医疗、司法等高风险领域的输出内容进行多重校验，有效降低单一模型的幻觉风险。此外，开发幻觉检测算法，对生成内容进行自动甄别与风险标注，可从源头遏制错误传播。据此情况，Gemini 发布时，推出了核查回答功能，用户使用此功能，Gemini 回答的每一个事实都进行一次 Google 搜索，然后通过对比来验证回答是否有据可查，甄别事件的真实。

4.2. 制度规制：设立“硬约束”与“软规范”治理体系

在法律层面，可借鉴欧盟提出的《人工智能法案》和欧洲议会通过的《AI 法案》的经验，对 AI 风险进行分级监督，形成通用型人工智能和基础模型的监管规则，要求模型的提供者应在对应数据库注册，对生成内容进行“数字水印+风险提示”双重标识，实现溯源可查。国家网信办联合国家发改委、教育部等七部门出台《生成式人工智能服务管理暂行办法》，明确需进一步细化幻觉责任界定，明确开发者、使用者在不同场景下的责任边界。

在行业治理层面，建立分领域幻觉评估标准。对政务、医疗等高危领域实施强制认证，要求模型通过特定场景的幻觉测试方可落地。推动行业协会建立黑名单制度，对故意投喂错误信息，利用 AI 幻觉传播不实数据的企业加大惩处力度，形成制度震慑。

4.3. 产业应用：推行“负责任的 AI”

AI 开发公司应具备企业自律，遵循“Safe-by-Design”原则，将幻觉缓解作为核心设计指标，而非事后补救。建立透明的模型卡、数据表，公开已知风险。

企业在引入 AI 时，必须重新设计工作流程，明确 AI 的使用规则和范围，嵌入“人类在环”的审核与监督环节，尤其在关键决策点确保人类行业专家的最终裁决权，而不过度依赖 AI。

4.4. 社会公众：实施“全民赋能”计划

用户素养提升是防范幻觉风险的最后防线。构建分层教育体系：对普通公众开展媒介素养教育，普及“AI 内容交叉验证”的使用习惯；对专业领域使用者进行专项培训，提高自身专业技术水平，掌握本行业的幻觉甄别方法。研究表明，经过系统培训的用户被幻觉误导的概率可显著降低。

5. 结论与展望

“AI 幻觉”并非技术发展的阶段性问题，而是伴随生成式 AI 全程的系统性挑战，它是一把“双刃剑”，其在效率提升、公共服务升级等方面的优势为社会发展注入强大动力，而伦理风险与安全隐患也对社会治理提出严峻挑战。技术本身并无善恶，关键在于人类如何驾驭，短时间内无法彻底解决幻觉问题，比起依赖 AI，我们更要依赖自己的学习和判断。未来，随着技术创新与制度完善，人工智能必将在人类社会发展中扮演更积极的角色，成为推动文明进步的重要力量。对其伴生的“幻觉”需要开发相对精准的量化评估工具，实现风险的动态监测，探索将社会伦理价值嵌入模型的技术路线，使“可靠性优先”成为 AI 的内生属性。唯有如此，才能在发挥 AI 技术潜力的同时，将幻觉风险控制在社会可承受范围，推动人工智能真正成为社会发展的赋能者。

参考文献

徐剑。人工智能为何会产生幻觉[J]。 *中国报业*，2025, (13):5.
李凌。构建大模型幻觉及其价值风险的预防治理体系[EB/OL]. *光明网*， 2025-04-11.
OpenAI, Georgia Institute of Technology. Why Language Models Hallucinate[R]. 2025.
徐延民。人工智能知识幻觉的生成逻辑与治理之道[N].*重庆科技报*，2025-08-07(012).
Casu M , Guarnera L , Caponnetto P , et al. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions[J]. Forensic Science International:Digital Investigation, 2024.

The Social Impact and Multidimensional Governance of AI Hallucinations

Yuantai Wu， Shiming Yuan， Xinyi liu

Abstract：Hallucination, as a characteristic of artificial intelligence (AI) technology, is one of the roots of various superpowers of AI, such as recognition and generation. With the development of generative AI, hallucination has also emerged as a technical defect, becoming a variable that affects social development. This article focuses on analyzing the generation mechanism of "AI hallucination", and explores its multi-dimensional impacts on the information ecosystem, social operation, technological development, and human progress. Ultimately, it constructs a collaborative governance system from four levels: technical optimization, institutional regulation, application layer, and social layer, providing a reference for promoting the safe and controllable development of AI technology.

Keywords： AI hallucinations; large model technology; generative AI; social risks; collaborative governance; multi-dimensional governance