Journalism and Communications 新闻传播科学, 2025, 13(8), 1515-1521
Published Online August 2025 in Hans. https://www.hanspub.org/journal/jc
https://doi.org/10.12677/jc.2025.138215

AI幻觉的多维解析：成因、现象及影响的全面阐释

王　蕊

河北师范大学新闻传播学院，河北　石家庄

收稿日期：2025年7月20日；录用日期：2025年8月13日；发布日期：2025年8月21日

摘　要

本文围绕AI幻觉展开多维解析，先界定其内涵为AI系统生成的不准确、虚构内容，本质是算法、数据与人类意图的错位。接着从技术层面的算法缺陷、数据质量瓶颈、模型可解释性矛盾，以及人类输入环节的提示词歧义、数据标注与反馈机制缺陷、人机交互认知依赖，分析生成机制。然后阐述其在内容生产、信息分发、商业服务等互联网场域的具体表现。最后探讨双重影响，消极方面有破坏信息秩序、消解机构公信力等，积极方面能激发创造性思维、推动技术进化等。指出需理性认知人工智能快速发展，从技术、制度、用户层面平衡应对，这是数字时代重新定义知识与智能的过程。

关键词

AI幻觉，算法黑箱，生成机制，双重影响，数据污染

A Multidimensional Analysis of AI Hallucinations: A Comprehensive Interpretation of Causes, Manifestations, and Impacts

Rui Wang

School of Journalism and Communication, Hebei Normal University, Shijiazhuang Hebei

Received: Jul. 20th, 2025; accepted: Aug. 13th, 2025; published: Aug. 21st, 2025

Abstract

This paper conducts a multi-dimensional analysis of AI hallucinations, first defining them as inaccurate

文章引用: 王蕊. AI 幻觉的多维解析: 成因、现象及影响的全面阐释[J]. 新闻传播科学, 2025, 13(8): 1515-1521.
DOI: 10.12677/jc.2025.138215

and fictional content generated by AI systems, with the essence being the misalignment between algorithms, data, and human intentions. It then analyzes the generation mechanism from the technical aspects of algorithmic flaws, data quality bottlenecks, and the contradiction of model interpretability, as well as the human input aspects of ambiguous prompt words, deficiencies in data annotation and feedback mechanisms, and cognitive reliance in human-machine interaction. It further elaborates on their specific manifestations in internet domains such as content production, information distribution, and commercial services. Finally, it discusses the dual impacts, with negative aspects including the disruption of information order and the erosion of institutional credibility, and positive aspects such as stimulating creative thinking and promoting technological evolution. It points out the need for a rational understanding of the rapid development of artificial intelligence and a balanced response from both technical, institutional and users' perspectives, which is a process of redefining knowledge and intelligence in the digital age.

Keywords

AI Hallucination, Intelligent Black Box, Generative Mechanism, Dual Influences, Data Pollution

1. AI 幻觉的内涵界定与本质特征

AI 幻觉是指 AI 系统，特别是基于深度学习的大型模型，在缺乏真实数据或有效约束的情况下，通过对其训练数据的过度泛化、不合理外推或创造性解释而生成的不准确、不真实、超现实甚至完全虚构的图像、文本、声音或视频内容[1]。今年 3 月 4 日，AI 评估组织 Vectara 公布的大语言模型(LLM)幻觉排行榜上，DEEPSEEK-R1 的幻觉率达到了 14.3%，这一数据远高于其 V2.5 版本的 2.4% [2]。这种现象并非偶然的技术故障，而是人工智能发展到特定阶段的必然产物，折射出当前技术体系与人类认知交互过程中的深层矛盾。从本质上看，AI 幻觉是算法逻辑、数据特征与人类意图三者错位的集中体现，其存在揭示了人工智能“智能”边界的局限性——即机器尚未具备真正意义上的事实判断与逻辑推理能力，更多是基于概率模型进行模式匹配与内容生成。

2. AI 幻觉的生成机制：技术缺陷与人为因素的双重作用

2.1. 技术发展的固有局限性

1. 算法模型的内在缺陷

深度学习模型，尤其是大型语言模型(LLM)，普遍采用 Transformer 架构，依赖自注意力机制处理序列数据。这种架构本质上是通过预测下一个 token 的概率分布来生成内容，而非基于对语义的真正理解。当模型面对训练数据中未覆盖的罕见场景或需要逻辑推理的问题时，极易因概率计算偏差产生幻觉。例如，GPT 系列模型在回答涉及复杂逻辑的问题时，可能会编造看似合理但事实错误的答案，这是因为模型更关注语言流畅性而非事实准确性。

2. 训练数据的质量瓶颈

数据是 AI 模型的“粮食”，但训练数据的局限性直接导致幻觉的产生。其一是数据偏差，若训练数据来自有偏见的数据源(如特定立场的新闻网站、非均衡的语料库)，模型会继承这种偏差并生成带有主观倾向的内容。例如，训练数据中缺乏多元文化视角，可能导致 AI 在生成涉及特定群体的内容时出现刻板

印象。其二是数据覆盖不足，面对海量知识领域，训练数据难以穷尽所有事实细节。当模型被问及冷僻领域和新兴问题时，由于数据库覆盖未能面面俱到，在生成内容时缺乏相关数据支撑，只能通过拼接和联想生成虚构信息。其三是数据噪声干扰，AI 大模型中的数据源于公开数据集、网络抓取数据、专业和授权数据、用户生成内容、合成和增强数据及众包数据。在用户生成内容与合成和增强数据环节，网络数据质量参差不齐，在网络数据中存在大量虚假信息、重复内容或错误标注，这些噪声数据会被模型学习并复制，在对已有错误数据的基础上随机扰动，如文本替换、图形拼接等，形成“错误传承”。例如，训练数据中若包含错误的历史事件时间线，AI 会将其作为“事实”传播。

3. 模型复杂度与可解释性的矛盾

随着模型参数量级从亿级向万亿级跨越(如 GPT-4 参数量超万亿)，其内部逻辑因高度复杂而形成“黑箱”系统，导致开发者难以追溯幻觉产生节点。这种黑箱特性源于大模型的“端到端”(E2E)架构，即从输入到输出，省略了从原始数据到最终结果可见性的过程[2]。同时大模型概率内生机制的训练设计缺陷，模型会倾向选择训练数据中高频表达方式，却无法自我纠正错误，进而引发错误级联放大。大规模模型在训练中易产生意料之外的“涌现能力”，伴随无逻辑概念拼接等不可预测的错误模式，加之模型在训练分布内任务表现优异但面对分布外输入时泛化能力失效，致使幻觉频发。例如 Open AI 的自动语音识别系统 Whisper 在医疗领域中将患者与医院的对话问诊过程音频转化为文字病例的过程中，2.6 万多份自动转录病例中，几乎每一份文件都存在编造和虚构的问题，这对患者健康和医疗系统产生严重负面影响。

2.2. 人类输入环节的系统性风险

1. 提示词设计的歧义性

用户与 AI 交互时使用的提示词若存在语义模糊或逻辑漏洞，会直接诱导幻觉。用户需求表述模糊时，如用户询问“推荐一本关于人工智能的好书”，若未指定具体领域(技术、伦理、历史等)，模型可能推荐不存在的书籍或错误作者。若用户提问中隐含假设误导，模型会基于该前提进行“合理”推导，生成错上加错的内容。例如，用户提问“如何看待外星人在 2020 年登陆地球事件”，模型可能虚构相关“证据”。

2. 数据标注与反馈机制的缺陷

在人工智能技术的应用进程中，数据标注与反馈机制的系统性缺陷成为诱发 AI 幻觉的重要成因。从数据标注环节来看，人类标注者的主观认知偏差在监督学习与强化学习中渗透至模型训练底层，不同标注者对语义边界的理解差异导致标注数据矛盾，如 Vectara 幻觉排行榜显示，DeepSeek-R1 因训练数据中中文历史语料覆盖率不足 35%，其在中文语境下对冷门历史事件的错误率较英文语境高出 27%，甚至生成“闸北区哪吒弄堂”等虚构内容。这种标注噪声进一步引发模型推理混淆，使 DeepSeek-R1 的幻觉率达 14.3%，显著高于 GPT-4o 的 1.8%。当标注数据携带群体偏见时，模型会通过“端到端”学习强化偏差，如图书馆 AI 推荐系统中 18%的性别歧视性推荐案例，即源于训练数据对特定职业的描述存在认知偏差[3]。反馈机制的滞后性与验证缺失则加剧了 AI 幻觉的传播与固化。用户对 AI 生成内容的过度依赖导致 83%的信息缺乏交叉验证，如美国律师使用 ChatGPT 生成的虚构法律案例提交法庭，若该内容被法律 AI 系统收录，将形成“伪先例”递归链。反馈链条的延迟效应使模型修正效率低下，用户对事实性错误的反馈平均滞后 48 小时，且仅 12%的错误被有效上报。更深层的风险在于，当 AI 输出符合用户预期但违背事实时，67%的用户会选择性接受信息，如 DeepSeek-R1 在医疗咨询中遗漏癌症诊断关键条件时，患者因“答案符合心理预期”而忽略验证，导致误诊风险提升 3 倍。这种“幻觉–反馈”的恶性循环，最终推动“奇幻社会”危机——当 AI 生成的虚构内容被纳入数字记忆，将系统性削弱社会对现实的认知

根基。“奇幻社会”指AI幻觉导致虚构内容渗透现实认知的社会状态[2]。

3. 人机交互中的认知依赖陷阱

人类在使用AI工具时，可能因过度信任而忽视对输出内容的验证。自动化偏见，用户倾向于认为AI生成的内容具有权威性，默认其准确性，从而减少主动核实行为。如记者使用AI撰写新闻时，可能直接采用包含幻觉的内容，造成虚假新闻传播。意图传递失真，用户难以将复杂的知识需求转化为AI可理解的形式，模型因误解意图而生成偏离目标的内容。例如，法律从业者要求AI“总结某案件的关键证据”，模型可能错误提取无关信息。

3. AI幻觉在互联网场域的具象化呈现

3.1. 内容生产领域的虚假信息泛滥

1. 新闻传播中的AI生成式幻觉

自动化写作工具在新闻生产中的普及带来了效率提升，但也成为幻觉的重灾区。DeepSeek等AI工具可能会在要求不清晰的情况下编造一些信息[4]。事件细节虚构，AI撰写的新闻可能编造不存在的采访对象、对话内容或事件经过。例如，某AI生成的财经新闻中，虚构了“某专家预测股市涨幅”的言论，导致市场误判。数据篡改与错误关联，在处理统计数据时，AI可能因算法错误或数据理解偏差，生成错误的趋势分析或因果关系。如将不相关的经济指标强行关联，得出“冰淇淋销量上升导致犯罪率增长”的荒谬结论。

2. 社交媒体中的信息污染

在社交媒体生态中，聊天机器人与内容生成AI的规模化应用正成为幻觉信息扩散的关键推手，其技术特性与传播机制共同加剧了后真相时代的认知危机。从虚假账号与水军内容的生成机制来看，AI依托批量注册的虚拟身份体系，能够高效产出伪造的用户体验、产品评价及热点评论，形成具有误导性的舆论浪潮。如杜骏飞在研究中揭示的，某品牌通过AI生成大量产品正面评论时，因模型训练数据的时间逻辑缺陷，出现“该产品2025年上市”的时间穿越幻觉——这种基于算法概率生成的内容，尽管存在事实性矛盾，却能借助社交媒体的流量逻辑快速渗透公众认知[2]。更具隐蔽性的风险在于谣言的智能化变种生产。AI通过抓取真实事件片段并填充虚构细节，能够制造“半真半假”的复合型谣言，其混淆现实与虚构的特性显著提升了信息辨识难度。例如将某地区普通事故篡改为“重大安全事件”时，AI不仅会生成逻辑自洽的文字描述，还能同步产出伪造的现场图片及视频素材，形成多模态的幻觉信息组合体。例如今年1月西藏发生6.8级地震，在社交平台上出现了一个戴着帽子的小孩被压在倒塌建筑物下面的图片，图片生动逼真引发大量网友关注和转发，经多方查证，发现这张图片和相关短视频是AI生成。在社交媒体的算法推荐机制下，此类谣言因契合用户情感偏好而获得流量加持，最终演变为杜骏飞所预警的“后真相递归”：每轮传播中14.3%的幻觉内容会被重新编码为新的训练数据，经多轮迭代后，虚假信息将逐步渗透数字记忆系统，削弱社会对现实的集体认知根基。

3.2. 信息分发系统的误导性推送

1. 智能推荐算法的偏差放大

推荐系统基于用户行为数据进行内容分发，但若数据中存在幻觉信息，会形成“错误循环”。智能算法推荐会加大错误信息的回音室效应，用户持续接收符合其偏好的内容，其中可能包含AI生成的虚假信息，导致认知偏差固化。AI可能误判某些幻觉内容为“高价值信息”(如标题党文章因点击率高被推荐)，导致优质内容被淹没，低质信息优先被推荐。如健康类平台推荐AI生成的“偏方治百病”文章，掩盖正规医学知识。

2. 搜索引擎的结果污染

AI 驱动的搜索引擎在处理复杂查询时，可能返回包含幻觉的搜索结果。知识图谱错误关联，搜索引擎的知识图谱若存在构建缺陷，会将不相关的概念错误链接。例如，搜索“爱因斯坦的化学贡献”时，AI 可能编造其在化学领域的虚构成就。实时信息误判，在处理突发新闻时，AI 可能基于不完整数据生成错误的事件描述，如错误报道事故伤亡人数或原因。

3.3. 商业与服务场景的信任危机

1. 电商与广告领域的虚假营销

AI 生成的广告内容可能包含夸大或虚构的产品功效。在性能指标方面伪造，如某电子产品的 AI 生成广告中，宣称“电池续航时间达 72 小时”，但实际产品仅支持 24 小时，这种幻觉信息构成消费欺诈。在用户评价方面造假，AI 批量生成的虚假用户评价中，可能出现不符合产品实际的使用体验描述，如“某护肤品使用后瞬间美白”，缺乏科学依据。

2. 客服与咨询系统的错误引导

智能客服在处理专业问题时，可能因知识局限产生幻觉，在医疗咨询领域，AI 医疗客服若误判症状，可能给出错误的治疗建议。例如，将胸痛简单归因于“肌肉劳损”，忽视心脏病的可能性。在法律领域的 AI 咨询工具可能错误解读法条，导致用户采取错误的法律行动，如错误计算诉讼时效。

4. AI 幻觉的双重影响：风险挑战与创新机遇

4.1. 消极影响：认知与社会层面的多重风险

1. 信息传播秩序的破坏

AI 生成的幻觉信息具有“类真实”特征，传统的事实核查手段难以快速识别，导致虚假信息以“高效率、大规模”的方式传播。例如，深度伪造(Deep fake)技术制作的视频，将虚构事件“可视化”，严重动摇信息可信度。大量幻觉信息涌入公共话语空间，使公众难以区分事实与虚构，加剧认知焦虑。如在疫情期间，AI 生成的“特效药”谣言引发抢购潮，干扰正常社会秩序。

2. 机构公信力的消解

随着我们迈入人工智能时代，大型语言模型的准确性问题在某些领域引发了人们的普遍担忧，在某些领域引起了人们的戒备心理[5]。媒体转发由 AI 虚构或编造的内容，会破坏自身的专业性与客观性。例如，某知名媒体使用 AI 撰写的新闻被发现存在多处事实错误，其品牌公信力大幅下降。AI 系统若在关键领域(如金融、医疗)持续产生幻觉，会削弱公众对技术的信任，阻碍人工智能的社会应用进程。如 AI 诊断系统多次误判癌症病例，患者可能拒绝接受 AI 辅助治疗。

3. 认知能力与思维模式的异化

过度依赖 AI 工具，人类可能减少对信息的独立验证与深度思考，当公众面对海量信息时，因认知负荷的增加，往往依赖算法对信息进行筛选而非主动验证，形成“认知惰性”[6]。例如，学生使用 AI 生成论文时，若不核查内容准确性，会丧失学术研究的批判性思维。长期接触 AI 幻觉信息，可能导致人类对“真实性”的判断标准降低，甚至将逻辑自洽的虚构内容视为“合理事实”。如某 AI 生成的历史小说被读者误认为真实历史记载。

4. 伦理与法律风险的凸显

AI 幻觉导致的损害事件中，难以明确开发者、使用者与 AI 系统的责任边界。例如，AI 生成的虚假广告引发消费者损失，广告主、平台与 AI 技术方可能互相推诿责任。幻觉信息可能包含伪造的个人信息或敏感数据，被用于网络诈骗、身份盗用等犯罪行为。如 AI 生成包含真实人物姓名与虚构经历的“黑

料”，损害他人名誉。

4.2. 积极影响：创新驱动与认知拓展的潜在价值

1. 创造性思维的激发

AI 幻觉产生的“意外输出”可能成为艺术创作的灵感来源。北京大学计算机学院教授黄铁军认为正是 AI 幻觉的产生才带来了创新的可能性。“幻觉”使人工智能创造性的体现，人类要想创造比自身更强的智能体，就不要降低 Aid 幻觉率，否则人工智能将与巨大的资源检索库无异[7]。例如，AI 生成的抽象画作或诗歌，因其非逻辑性和独特性，为艺术家提供全新创作视角。如 Google 的 Deep Dream 系统通过算法幻觉生成的艺术作品，开创了“计算艺术”的新流派。当 AI 在解决问题时产生幻觉，可能意外突破常规思维框架，提出创新性方案。例如，某 AI 在设计电路时，因算法偏差生成了非传统但高效的电路结构，为工程设计提供新思路。

2. 认知边界的探索工具

AI 幻觉反映了当前机器学习模型的缺陷，同时也映射出人类认知的局限性——模型的“错误模式”可能与人类的认知偏差(如确认偏误、联想跳跃)存在相似性。通过研究 AI 幻觉，认知科学可更深入理解人类思维机制。例如，AI 在逻辑推理中的错误模式，与人类在相似任务中的常见失误高度吻合，为认知心理学研究提供数据支撑。AI 无法回答或产生幻觉的问题，往往揭示了人类知识体系中的空白领域。例如，AI 在处理量子物理与哲学交叉问题时的混乱输出，提示该领域需要更多跨学科研究。

3. 技术进化的反向驱动力

AI 幻觉为技术开发者提供了明确的改进方向，推动模型向更高准确性发展。例如，针对幻觉问题，研究者提出“事实性增强”技术，通过引入外部知识库(如维基百科)验证生成内容的真实性，降低幻觉发生率。在 AI 伦理问题层面上，幻觉问题促使行业重视 AI 伦理建设，推动透明度、可解释性等标准的制定。如欧盟的《人工智能法案》明确要求高风险 AI 系统需具备事实核查机制，从法律层面倒逼技术合规。

4. 应用场景的创新拓展

在广告设计、游戏开发等领域，AI 幻觉可被“可控利用”，生成新颖的创意方案。例如，游戏公司利用 AI 生成随机剧情分支，其中部分“幻觉剧情”因独特性受到玩家喜爱。通过分析 AI 幻觉案例，学生可学习如何辨别信息真伪，提升信息辨识的批判性思维能力。如教育平台设计“AI 幻觉识别”课程，通过对比真实与虚构内容，增强学习者的信息素养。

5. 应对之策

5.1. 技术层面

数据是 AI 大模型生成的源头活水，想要修正 AI 幻觉的漏洞必须修正原有数据库中的错误信息，加强数据爬取审查。2023 年国家网信办等部门颁布的《生成式人工智能服务暂行管理办法》中明确要求使用合法的数据和模型[8]。针对 AI 生成虚假信息问题，构建跨机构数据协同机制，通过区块链技术追踪数据来源，实现信息机构、学术数据库、政府公开数据等多源信息的跨域共享，以提升数据的真实性和可溯源性[7]。同时引入其他 AI 模型，以技术监测技术的方式，实现人工智能之间的互相监测和纠正。

5.2. 审核层面

目前关于生成式人工智能技术的规定已明确要求提供者应当对图片、视频等生成内容进行标识[8]。但随着人工智能生成内容越来越发达，生成内容与真实内容更难区分，这种单一要求提供者标识的审核制度难以应对更加复杂的应用场景。审核不仅要从内容提供者入手，更要从平台入手。平台应建立人工

智能和专家联合的审核小组，一方面通过优化 AI 算法，全面应对各种复杂的应用场景，应利用人工的方式对其进行辨别、审查，尤其是在医疗、金融等专业领域，要发挥专业从业者的作用，保障内容的专业性和可靠性。AI 生成式内容的监管还需多方协作，政府应主导风险分级与制度规范，企业承担合规与自查责任，第三方机构参与检测与认证，建设跨平台共享机制与动态监测系统，实现跨领域协同治理，推动 AI 技术在安全边界内健康发展[6]。

5.3. 用户层面

用户作为 AI 内容的接收者与传播者，其媒介素质的提升是应对 AI 幻觉的重要一环。在信息爆炸的环境中，用户需主动培养对 AI 生成内容的辨别意识：例如，面对可疑信息时，通过交叉验证(如比对权威信源、核查数据出处)判断真实性；关注内容中的逻辑漏洞或异常细节(如时间矛盾、事实错误)，警惕“看似合理却虚构”的 AI 幻觉特征。同时，用户需掌握与大模型交互的提问技巧，以降低受 AI 幻觉影响的风险：首先应优化提问方式，向 AI 提供更明确、详细的背景信息与上下文语境，为其生成内容锚定更精准的方向；其次可要求大模型分批输出结果，例如先列提纲再分段呈现内容，随后逐段进行审核，通过分步验证减少整体偏差；第三，在获取答案后，可进一步“追问”大模型，要求其标明内容的文献来源或原文链接[9]。这种个体层面的能力提升，既能有效减少对 AI 幻觉内容的误信与传播，又能与技术治理、平台审核形成协同，共同构建更立体的风险防控网络。

6. 结语

在正视与驾驭中寻求平衡 AI 幻觉作为人工智能发展的伴生现象，既暴露了技术的局限性，也蕴含着创新的可能性。面对这一复杂议题，简单的“抵制”或“放任”均非良策。我们需要以理性态度认知其生成机制，在技术层面推动模型优化与事实核查体系建设，在制度层面完善监管框架与伦理规范，同时在应用层面探索幻觉的创造性转化。唯有如此，才能在人工智能的浪潮中，既守护信息真实性的底线，又释放技术创新的红利，实现人机协同的良性发展。从更深远的视角看，对 AI 幻觉的理解与应对，本质上是人类在数字时代重新定义“知识”“真实”与“智能”的过程，这一过程将深刻影响未来社会的认知格局与技术伦理。

参考文献

[1] Hatem, R., Simmons, B., Thornton, J.E. (2023) A Call to Address AI “Hallucinations” and How Healthcare Professionals Can Mitigate Their Risks. *Cureus*, **15**, e44720. https://doi.org/10.7759/cureus.44720

[2] 杜骏飞. 奇幻社会的来临——DeepSeek 幻觉与后真相递归[J]. 探索与争鸣, 2025(3): 11-14+177.

[3] 裴宏娇, 郭瑞宇, 杨志和. 图书馆 AI 语言生成中的信息幻觉: 信息素养视角下的风险规避策略[J]. 图书情报导刊, 2024, 9(12): 36-43+53.

[4] 张聪聪. DeepSeek 加速出版变革但需警惕 AI 幻觉[N]. 中国出版传媒商报, 2025-03-14(010).

[5] 冯岩, 奥利弗·鲍恩. 当 AI 出现幻觉时[J]. 世界科学, 2024(3): 28-30.

[6] 李佳凯, 李一. AI 幻觉驱动下信息生态的信任危机及其治理路径[J]. 融媒, 2025(7): 29-36.

[7] 姬晓婷. 北京大学计算机学院教授黄铁军: 不能简单地将 AI 幻觉“一棒子打死” [N]. 中国电子报, 2024-05-07(007).

[8] 中国政府网. 生成式人工智能服务管理暂行办法[EB/OL]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm, 2023-07-10.

[9] 任翀. 避免“AI 幻觉”, 怎么提问很重要[N]. 解放日报, 2025-02-21(006).