技术与文化：生成式人工智能幻觉的认知僭越及其治理转型 *

曹咏萍

[摘　要] 生成式人工智能的幻觉现象不仅是一种技术性误差，更是一种深层次的认知僭越，既挑战了人类知识的边界与验证机制，又动摇了社会信任与知识合法性基础。现有治理路径多集中于技术优化与制度规制，却忽视了幻觉在文化语境中的生成与扩散机制。本文从哲学与文化双重视角出发，将生成式人工智能幻觉视为技术逻辑与文化建构交织的产物，并据此提出“认知僭越”这一核心框架。该框架旨在揭示生成式人工智能幻觉如何通过多种方式获得其合法性，具体包括技术权威的预设、信息过载环境下的思维惰性，以及文化叙事的固有偏向。在此基础上，本文主张生成式人工智能幻觉的治理范式应从单一纠错转向技术、文化与制度的三维协同：技术上增强可验证性与透明度，文化上培育批判性思维与算法素养，制度上构建责任明晰与多元共治的机制；治理的目标并非彻底消除幻觉，而是通过认知秩序的重建将幻觉转化为推动人机共生与主体性再确立的契机。

[关键词] 生成式人工智能　幻觉　认知僭越　文化自觉　治理转型

[中图分类号] B829　　[文献标志码] A　　[文章编号] 1009-8461（2025）11-056-14

一、引言及文献综述

以大型语言模型（large language models, LLMs）为技术基础的生成式人工智能（generative artificial intelligence, GAI）在诸多领域展现出强大能力，它的迅猛发展正重塑知识生产与社会认知的基本范式。然而，其固有的“幻觉”（hallucination）问题也日益凸显。生成式人工智能幻觉（以下简称“AI幻觉”）通常指大语言模型生成的文本不忠于信息源或者与现实世界的事实不符（刘泽垣等，2025）。与传统错误不同，AI幻觉不仅关乎技术准确性，更在深层次上挑战了人类的认知边界与知识秩序，构成一种系统性的认知僭越（cognitive transgression），侵蚀人类判断真伪和评估可信度的认知根基，从而动摇社会信任赖以存在基石。

当前学界对AI幻觉的治理研究主要遵循三条路径：一是技术优化路径，侧重于通过模型架

* 作者简介：曹咏萍，广东省社会科学院哲学与宗教研究所助理研究员。

构优化与创新、数据清洗与提示工程等技术手段降低幻觉频率（刘泽垣等，2025）；二是制度规制路径，聚焦于通过立法、标准与监管框架（如欧盟《人工智能法案》）来界定责任、管控社会风险（许可，2025）；三是初显的治理转型路径，开始倡导跨学科的多维治理框架（刘永谋和孙瑞璇，2025）。现有研究虽然提供了一些重要的研究基础，但仍局限于“错误修补”范式，未能充分阐释 AI 幻觉对知识合法性的深层冲击，尤其缺乏对文化语境中 AI 幻觉生成与扩散机制的深入分析，同时也缺乏从哲学层面反思其作为“认知僭越”的深远意义。尽管有研究开始跳出纯技术范式，尝试从技术演化的过程性视角剖析 AI 幻觉，如胡泳和王昱昊（2025）基于技术过程论提出 AI 幻觉的“价值负荷”是一个动态过程，其伴随技术从构思、发明到产业化的演进，多元行动者共同参与并博弈，最终形成复杂的价值共赋现象，为理解 AI 幻觉的社会建构性提供了重要启发，但其分析未充分揭示这一“价值负荷”过程如何动摇了人类的认知验证机制与知识合法性基础。此外，亦有学者从技术哲学与知识生产范式变革的视角进行深度剖析。王鸿宇和蓝江（2025）指出，生成式人工智能推动了知识生产从“物性”向“无人化”的转变，其幻觉现象乃是“外主体”进行符号自我增殖的逻辑必然，并系统揭示了此举对知识合法性基础构成的结构性风险与认识论危机，为理解 AI 幻觉如何动摇知识生产的传统根基提供了宏大的理论视野。不过，该分析的重心在于技术形态演进史与知识权威的宏观转移，对于两个关键维度仍缺乏深入探讨：一是 AI 幻觉在具体文化语境中的生成与扩散机制，二是其作为“认知僭越”对个体认知框架、社会信任机制的微观作用路径。此外，刘泽垣等（2025）的技术性综述从实证层面梳理了 AI 幻觉的成因、评估与缓解路径，为本文提供了技术理论依据。

基于上述研究的启发与仍待深入的解释空间，本文尝试将 AI 幻觉置于技术逻辑与文化建构交织互构的整体视野中加以考察。本文提出“认知僭越”这一核心理论框架，旨在系统阐释 AI 幻觉如何凭借其技术拟真性与文化合法性表象，动摇传统知识验证的逻辑基础，进而影响社会认知秩序的稳定性。需要说明的是，本文虽以技术与文化为主线聚焦 AI 幻觉的生成机理，但因制度在技术与文化互动中扮演着关键的中介与固化角色，所以构建的治理方案明确包含制度维度。遵循这一理路，本文首先剖析 AI 幻觉引发的认知僭越及其微观作用机制，随后解构其技术与文化交织的生成逻辑及合法化过程，最终致力于构建一个融技术可验证性、制度可追溯性与文化可反思性于一体的三维协同治理范式，以期为人机共生环境中认知秩序的重建和主体性价值的守护提供理论参考与路径探索。

二、生成式人工智能幻觉的认知僭越

生成式人工智能的幻觉现象，不仅限于技术层面的输出偏差，而且还是一种深嵌于认知逻辑与知识结构之中的系统性“僭越”。它跨越了工具性辅助的边界，介入甚至重构人类的理解方式与判断机制，进而对个体认知、知识合法性以及社会信任产生多层次的冲击。

（一）个体认知的边界模糊与验证困境

生成式人工智能的“幻觉”现象，首先在个体层面引发了深刻的认知挑战。与传统的信息失真不同，AI 幻觉并非偶然误差，而是源于其概率生成机制的内在特性。它通过高度拟真的语言

形式，系统地模糊了真实与虚构的边界，从而动摇了人类长期以来依赖的认知判断基础。

在传统认知模式下，人类依赖常识、直觉与逻辑推理来辨别真伪。然而，生成式 AI 输出的文本具有极强的语义连贯性与风格一致性，这种形式上的完善性极易对用户的认知惯性形成误导。AI 正是通过高度成熟的自然语言处理（natural language processing, NLP）与生成能力构建出这种极具迷惑性的权威外观。这种构建 AI 输出权威性的自然语言在知识生产中扮演着双重角色：它既是人机交互不可或缺的桥梁，却又可能化身为知识权威的隐形枷锁（王鸿宇和蓝江，2025）。这种通过自然语言所建立的权威性，与人类的认知心理机制相互作用，进一步放大了 AI 幻觉的影响。当 AI 以逻辑严密、表述流畅的自然语言填补知识空白时，用户常常因文本表面上的完整与确定而放松警惕，不再进行深入的追问与验证。特别是当这些文本以非常礼貌甚至有些“谄媚”的方式呈现出来时，人类对于 AI 幻觉识别的敏感性大为降低（Pak et al.,2024）。这种认知惰性如果在医疗、法律、教育等高风险场景中出现，可能会引发实质性的判断失误与行动后果。

不仅如此，AI 幻觉在对人类认知惯性进行误导的过程中还具有瓦解传统知识验证基本路径的风险。人类知识秩序建立在证据可追溯性与系统化审核机制之上，如学术引用、事实核查与同行评议等制度性安排。然而，生成式 AI 基于统计规律而非事实指涉的文本生成，其输出缺乏内在的事实锚点与明确的出处指引。普通用户缺乏验证工具和辨别能力，难以区分 AI 生成内容中的隐性误差；即便对专业者来说，甄别这些隐性误差也需付出极高的时间与认知成本。这种“证据缺席”的状况，使建立于现代性基础上的知识验证体系陷入深层困境。

此外，当代信息环境的特征加剧了验证机制的失效。在信息过载与内容碎片化的背景下，用户逐渐形成以“认知节约”为导向的信息处理策略，更倾向于接受便捷、现成且符合预期的答案，而非主动开展深度求证（刘明君，2025）。AI 幻觉恰以权威性、即时性、完整性与礼貌性的外在形态，为用户提供了认知捷径，从而催生了普遍的默认接受的心理图景，这是人类“心智成本最小化”的体现（刘永芳等，2025）。这种心理机制不仅弱化了个体的批判性思维，更使 AI 幻觉得以从个体误判扩散至群体认知偏差，最终对社会层面的知识共识与信任结构构成系统化威胁（黄立赫，2025）。

（二）知识合法性与社会信任机制的动摇

AI 幻觉的影响远不止于个体认知层面的信息偏差，还进一步渗透至知识合法性建构的核心环节与社会信任机制的底层逻辑。人类在数百年的知识实践中，逐步形成了对学术术语的严谨性、数据引证的溯源性、专业表述的逻辑性等特定符号形式的自觉信任。这些符号并非孤立的语言工具，而是知识生产过程中的外在载体，被普遍视为知识有效性与权威性的显性标志。例如，学术论文中引用权威文献的表述范式，背后是对前人研究成果的认可与学术传承的尊重；医疗诊断报告中“基于某某机构检测数据”的表述，承载的是医学领域对实证依据的严格要求。然而，基于大语言模型的生成式 AI 打破了这一符号与事实的绑定关系。它通过对海量文本数据的学习，精准复现了学术论文、专业报告、权威解读的语言风格与符号形式，却可能在算法生成过程中脱离符号本应指向的事实基础。生成式 AI 或是虚构不存在的文献引用（如 ChatGPT 曾生成伪造的学术参考文献）（Cheng et al.，2025），或是编造无数据支撑的专业结论，最终输出一种

“语法完备、格式规范但事实空缺、逻辑断裂”的伪知识。这种伪知识的隐蔽性的影响远超语言粗糙、逻辑混乱的传统谣言，因完美契合专业符号体系而具备了“权威外衣”下的欺骗性。

不仅如此，这一认知偏差过程还被现代社会中根深蒂固的“技术权威预设”进一步放大。自工业革命以来，技术被赋予“中立、客观、高效”的象征意义：蒸汽机的发明推动生产力飞跃，让人们信任机械技术的可靠性；计算机的普及实现信息高效处理，让人们默认数字技术的准确性。这种长期形成的技术信任惯性，在生成式 AI 这一前沿技术上得到了极致延续：作为“能理解、会创作”的智能系统，生成式 AI 天然被赋予了“超越人类局限”的权威光环。用户在面对其生成内容时，往往不自觉地降低质疑阈值：在法律咨询中，默认 AI 给出的条款解读符合法律条文；在教育辅导中，轻信 AI 提供的知识点解析等同于教材权威表述。当生成式 AI 被片面赋予认知权威的角色，其幻觉便不再局限于个体认知偏差，而是在群体互动中形成“合法性扩散”：某一领域的 AI 伪知识经专业人士误判后，可能通过行业报告、媒体报道进一步传播；在教育、医疗、专业咨询等对知识准确性要求极高的领域，AI 生成内容甚至对传统专家系统形成局部性、辅助性的替代性冲击。

AI 幻觉迫使我们必须重新审视“真理 – 权威 – 信任”的经典三角关系。在传统框架中，真理通过证据和逻辑获得权威，进而赢得社会信任；而在生成式 AI 所塑造的语境下，权威很大程度上来自技术的符号资本与拟真效能，信任则源于对技术系统的无意识服从，真理反而在这一过程中被悬置甚至边缘化。这一转变不仅意味着知识合法性基础的动摇，更揭示出一个严峻议题：当知识的权威被技术拟像所部分接管，我们应如何在新的认知生态中重建真正值得信赖的真理机制与知识秩序？

（三）人机共生关系中的认知僭越与主体性危机

生成式人工智能的深度融合正推动人机关系从“工具使用”迈向“认知共生”，然而在这一转型过程中，幻觉现象凸显了其潜在的认知僭越本质，即人工智能系统逾越其应有的工具性边界，介入甚至重塑人类的知识判断与认知权威，进而引发深层次的主体性危机。

这一僭越首先体现为人类认知活动的大规模外包化与随之而来的依赖症（王鸿宇和蓝江，2025）。这种依赖症的形成，根源在于生成式 AI 通过其高度拟真的交互能力，营造了一种“主体间性”的错觉。殷杰（2024）所描述的 AI 在对话中呈现出的目标导向性和自主性“交互主体性”在此恰恰成为认知僭越的桥梁：用户不自觉地进入“对话伙伴”而非“工具使用者”的关系模式，从而放松了批判性警惕。然而，必须清醒认识到，这种“主体性”只是一种由算法驱动的表象（刘永谋和孙瑞璇，2025），其本质仍是统计概率的运算，而非真正的意向性和意识。因此，这场僭越实际上是一个强大的工具成功地模仿了主体的外在行为，诱使真正的人类主体主动让渡了自身的认知权威。周研和沈天健（2024）也持有类似的观点，认为生成式人工智能“依然占据着人机互动的高位，看似让渡了交流的权利，其实仍隐秘地向用户主体施加着技术权威，粗暴地通过撒播的方式回应用户的话语或指令”。这表明，技术权威不仅是一种宏观预设，更是一种在微观交互中得以实践的、具有支配性的传播结构，它系统地麻痹了用户的批判性神经，为用户接受幻觉铺平了道路。随着生成式 AI 广泛应用于教育、研究、写作与决策支持等多个场景，个体逐渐将信息检索、内容生成乃至逻辑推导等核心认知任务交由机器完成。尽管这种做法显著

提升了效率，却也导致认知能力的外部化与内在批判意识的弱化。正如美国学者尼尔·波斯曼（2007）在《技术垄断：文化向技术投降》中所警示的，当技术不仅替代人类的肢体与感官、更进一步侵入了思维与解释的领域时，人便面临从“认知主体”滑向“被动接受者”的风险。生成式人工智能正在这一临界点上发挥作用：它不仅回应人类的提问，也重新定义了提问的方式与回答的框架，从而潜移默化地塑造了人类的认知模式与知识结构。

因此，认知僭越的根本问题不在于机器“犯错误”，而在于人类在技术逻辑面前逐渐放弃了批判、审辨与最终判断的权力。AI 以概率模型生成内容，却被赋予认知权威；人类默认其输出的有效性，不再回归自身作为知识审查者与意义赋予者的根本角色。这种权力关系的重构，标志着人机共生并未真正实现“共生”，而在某些维度上演变为一种“认知替代”。马丁·海德格尔（2020）对技术本质的论断在此显得尤为紧迫：技术从来不只是工具，而是一种根本性的“解蔽”方式，它塑造着我们理解世界和遭遇真理的基本模式。生成式人工智能正通过其语言建模与知识再造能力，深刻介入人类“在世界之中存在”的方式。倘若缺乏自觉的批判与治理，其所定义的“真实”可能凌驾于人类自身的体认与反思之上。米歇尔·福柯（2019）关于“权力 – 知识”共生关系的论述，为此提供了进一步的解释框架：权力通过知识确立自身，知识则承载并再生产权力。在生成式人工智能所塑造的认知秩序中，权力并未消失，而是嵌入算法的设计、训练数据的选取与模型生成的结果之中，并通过认知僭越获得合法性。因此，重构人类在技术条件下的主体性，不仅是一个技术或伦理问题，更是一项关乎认知正义与知识民主的政治性议题。

三、生成式人工智能幻觉的技术逻辑与文化意涵

生成式人工智能中的“幻觉”现象，远非一种可被单纯技术化归因或修正的系统偏差，而是深植于其概率性认知架构与文化语境相互构建的复杂产物。它既反映出基于统计建模的机器认知模式与人类求真传统之间的断裂，也揭示了技术系统在知识生成过程中对既有文化权力结构的隐性承接与强化。要真正理解 AI 幻觉的生成机制与深层意涵，就必须超越“模型 – 数据”二元框架，转而审视其如何在技术逻辑与文化嵌入的双重作用下被共同塑造、成为一种同时关乎算法结构和社会意义的复合性现象。

（一）技术层面的结构性局限

生成式 AI 的核心机制是基于 Transformer 架构的自回归预测，其基本任务不是传递真理，而是根据上下文语境生成概率最高的词序列（Vaswani et al.，2017）。在这一过程中，大语言模型并未建立对外部世界的表征关系，而是通过高维参数空间中的向量运算模拟语义关联。这种运作机制决定了生成式 AI 的输出本质是一种“统计拟真”，而非事实性断言。例如，当大语言模型在缺乏确切知识的情况下仍生成看似完备的答复时，它并非在撒谎，而是在执行其预设的语义补全任务。AI 幻觉因而不是系统故障，而是大语言模型作为一种概率机器的逻辑必然。

大语言模型所依赖的训练语料库并非客观中立的“世界镜像”，而是来自人类既有的文本记录，内含大量非均匀分布的文化假设、历史偏见与叙事结构。诸如英语语料的支配性比例、科技

文献的西方中心主义、性别与种族在描述中的不对称等，都会被算法无损学习并加以放大。在“预测下一个词”的生成过程中，大语言模型并非简单地复现已有偏见，而是将其系统性地整合并融入高度连贯的语义流之中，从而构建出一种在技术上高度自洽、表达流畅的偏见形态，形成所谓的“系统化偏见”或“结构性偏见”（赵越等，2024）。它不再以明显失真的方式呈现，而是隐藏在合乎语法、引证规范且逻辑自洽的文本内部。神经网络高达数百亿级的参数规模与多层次非线性变换，使其决策路径变得高度不透明，当大语言模型产生幻觉时，即便具备某领域内相关知识的用户也难以追溯其形成原因。系统固有的模糊性并非偶然缺陷，而恰恰源于其构建基础，即概率性生成机制在追求语义连贯性的过程中牺牲了对外部指涉真实性的保证。AI 幻觉因而成为技术理性中“工具性”压倒“理解性”的集中体现，也折射出智能系统在认知权威性与解释力之间的深刻断裂。

（二）文化层面的嵌入、再生产与合法化机制

AI 幻觉在文化维度上揭示了一种深层的符号运作机制。它不仅再现了训练数据中存在的文化偏向，更通过生成过程的自然化与权威化，将特定的文化框架塑造为认知上的“默认真实”，从而实现文化意义上的嵌入、再生产与最终合法化。文化嵌入远非表面层次的文本复制，而是一种深度的认知图景建构。

正如前文所讨论过的生成式 AI 中的偏见问题，大语言模型通过海量语料所学习的不仅是词汇与句法，更是一整套隐匿的文化想象与价值排序，其中包括西方中心的知识体系、男性主导的叙事传统、殖民历史观所隐含的空间与时间秩序等等，它们被转化为参数空间中的概率关系，悄然引导文本的生成方向。这些选择并非出自某种主观意图，而是模型在统计规律驱动下对主流文化模式的无声巩固，其后果是将某些文化的经验普遍化，而将其他文化表达推至边缘甚至不可见的领域。

大语言模型通过高度流畅、结构严谨的文本生成，将这些文化预设重新包装为看似中立且客观的“事实”。它不再仅仅复述已有的偏见，而是主动为其构建逻辑自洽的语义环境，生成虚构但可信的引文、模拟严谨的学术论证甚至拼合具有表面合理性的历史线索。在这一过程中，偏见不再显露为可辨别的谬误，而是深植于连贯的叙述肌理之中，被包裹于技术权威的外衣之下，成为一种更难以反思和批判的“算法文化无意识”。而这种无意识，最终在与用户的交互仪式中获得广泛认可与合法性。生成式 AI 凭借其响应迅速、文本完备、语气肯定的输出特性，逐渐让用户被培育出一种习惯性依赖，当用户反复接收到清晰、流畅且符合其文化预期的内容时，往往不再追溯生成结果背后的视角局限、历史条件与文化前提。大语言模型所输出的文化预设，因此被默认为理所当然的真实，人们无形中接纳了其中所承载的价值排序与符号秩序。生成式 AI 由此成为一种新型文化代理者——它既系统性地重塑了人们对“何谓真实”“何谓合理”的感知框架，也深刻参与了社会共识的生成与固化。AI 幻觉因此远超技术缺陷的范畴，成为一种值得深入批判的文化现象：它涉及意义的生产、叙事的权力与认知的权威，并在本质上重构了当代文化经验的生成方式。

（三）技术与文化互构的 AI 幻觉生成逻辑

AI 幻觉源于技术架构与文化语境的深度互构，它既非客观世界的镜像，亦非价值中立的符

号集合。在操作层面，这种互构体现为一种循环增强的生成逻辑。模型从语料中习得的不仅是语法，更是一整套隐匿的文化假设与认知框架，哪些观点更常见、哪些叙事更“合理”、谁的声音更值得被倾听，皆被编码为参数空间中的概率权重。而当模型基于这些权重生成文本时，它并非机械复现原有偏见，而是执行一种创造性的扭曲：碎片化的文化素材被重新组织为逻辑自洽的叙事，隐含的价值判断被转译为表面客观的陈述，有争议的立场则被配以逼真的佐证和权威的修辞形式。正是在这一过程中，偏见获得了新的存在形态：它不再容易被识别和批判，而是内嵌于模型输出的每一个环节，成为一种技术上合理、表达上流畅的“系统化幻觉”。

社会认知习惯与大语言模型生成特性之间的互动，进一步巩固了这一循环。用户倾向于信任那些结构清晰、语气笃定且符合自身文化预期的文本输出，不自觉地将生成内容纳入个人的认知参考框架。这种“认知舒适区”的建立，使得大语言模型再生产的文化假设得以绕过理性审辨，直接塑造用户对世界的理解（黄立赫，2025）。人机交互因而成为一种文化实践：技术在不知不觉中引导认知，而社会的接纳与信任则反向赋予技术输出以合法性，使特定的文化视角不断获得强化和自然化。

由此，我们可以清晰地看到在 AI 生成逻辑中内含着技术与文化之间的深度互构关系：大语言模型所输出的每一个句子，既是数学运算的结果，也是文化权力的具现；它既由历史中的数据所决定，也正在悄然书写新的历史。在这一意义上，生成式 AI 的幻觉生成了一个新的文化场域：旧有的不平等在此被重新编码，而新的认知秩序也正在人与机器的持续交互中逐渐成形。

（四）多元主体的实践：技术 – 文化互构的现实微观基础

技术逻辑与文化语境的互构关系，并非悬浮于理论层面的抽象推演，而是通过多元主体在技术生命周期中的具体实践得以具象化与固化。胡泳和王昱昊（2025）基于技术过程论的研究揭示了这一过程的微观机制：在技术构思阶段，设计者对模型创造力的偏好（如追求新颖性而非绝对准确）作为一种价值前提被嵌入技术的原始架构之中；到技术产业化阶段，商业资本通过精准调控模型参数（如温度设置）以实现流量汲取或舆论引导的目的，用户则通过有意无意的提示词工程，持续引导模型生成契合其自身认知偏好与价值立场的内容。此时，AI 幻觉不再只是一个技术问题，更像是一面棱镜，折射出社会价值与权力关系如何通过技术系统的中转与再生产，获得了新的表达形式与权威性。

这种多元主体的实践共同编织了一张精密的技术 – 文化互构之网。设计者的初始理念、资本的运行逻辑、用户的情境化互动，并非孤立发挥作用，而是在算法这个关键中介的转译下，彼此交织、互渗，甚至相互催化。训练数据中的文化偏见由此不再处于静态之中，而是被各类主体的实践持续激活、放大并赋予其新的现时相关性。在这个过程中，技术系统不仅仅是文化偏见的传递者，更成为偏见再生产的加速器，通过其强大的生成能力将局部价值观转化为具有普遍性表象的知识输出。这种转化之所以可能，正是因为多元主体在技术发展的各个阶段都不自觉地参与了“价值负荷”的过程，而技术本身的自动化与规模化特性又使得这种负荷过程变得隐蔽且高效。最终，幻觉的生成超越了简单的统计概率问题，成为一个社会技术建构过程。它揭示了所谓技术理性背后运行着一套由多元主体共同书写的、充满价值判断的社会理性，而这正是技术 – 文化

互构最为深刻的微观基础。这种微观基础的揭示，不仅让我们认识到幻觉产生的复杂性，同时也为我们理解如何在这种复杂性中寻求治理之道提供关键切入点。

四、生成式人工智能幻觉的多维治理

生成式 AI 幻觉是一种由多行动者参与建构的社会技术现象，不仅根植于技术与文化的深层互构，也是由设计者、资本、用户等多元主体在技术生命周期中的具体实践与互动所共同塑造。它既体现了概率模型的内在特征，也折射出社会知识生产中的结构性偏见。这些属性决定了对其治理需超越单纯技术纠错的范畴，要成为一项关涉如何协调不同主体的认知实践、重建知识权威分配机制的系统性工程。

（一）现有治理路径的局限性

当前主流的治理方法主要围绕“修正 – 控制”范式展开。无论是算法优化、数据清洗，还是实时核查与安全护栏机制，其核心逻辑仍是试图通过提升大语言模型的“准确性”来抑制幻觉发生的概率。然而，这类方法存在很大局限：基于概率生成的模型并不生产事实判断，而是在执行语义补全，这意味着 AI 幻觉并非系统故障，而是其固有逻辑的必然呈现；若执着于彻底消除 AI 幻觉，反而会削弱模型的泛化能力与创新潜力（殷杰，2024），是一种治标不治本的做法。那些具有高度连贯性、表面可信度极高的幻觉文本，正因为其形式上的合规性与权威性，更容易规避传统的内容审核机制，特别是在医疗诊断、司法判决、公共政策制定等高风险的决策场景中，AI 幻觉即便出现的概率低，也可能引发重大后果（刘泽垣等，2025）。由此可见，若现行治理模式仅聚焦于统计幻觉的发生频率，便会忽略一个重要事实：AI 幻觉借由技术系统营造的合理外观正悄然重塑人类认知与知识生产的底层逻辑。

在法律与政策层面，现有治理框架如欧盟《人工智能法案》虽以基于风险的分级监管为核心逻辑，但其治理模式仍很大程度上依赖于可预见的风险分类、事前合规评估及清晰的权责界定（崔星璐和姚长青，2025）。然而，AI 幻觉所带来的主要挑战恰恰在于它通过虚构引证、模拟学术论证和构建逻辑自洽的虚假叙述，系统性地动摇了传统上区分“事实”与“虚构”的认知边界，而这一边界正是许多既有法律与认知监管体系的基础，这种模糊直接导致了责任追溯的困境（刘永谋和孙瑞璇，2025）。当 AI 生成的内容造成危害时，其责任在开发者、部署者、用户之间高度分散和模糊，开发者常以“模型‘黑箱’”和“技术固有局限”为由抗辩，使用者则归咎于模型设计缺陷。这种局面使得建立在清晰因果关系与主观过错之上的传统法律归责模式难以有效运作，暴露了现有制度在应对技术自主性带来的责任真空时的结构性不足。比如，当大语言模型生成的法律判决书援引根本不存在的判例却能够通过司法审核时，这不仅揭示了算法生成机制的内在缺陷，也暴露出建立在确定性认知框架之上的专业验证系统的结构性漏洞。若法律制度依旧主要侧重事后追责与合规审查，却迟迟未发展出有效机制去介入 AI 幻觉从被误认、被接受到最终嵌入知识体系的扩散链条，便难以应对随之而来的认知框架松动以及这种松动向社会领域的蔓延态势。

（二）走向多维协同的治理范式

面对现有治理路径在应对幻觉生成的社会技术复杂性方面存在的根本局限，我们亟须构建一个融技术修复、制度约束与文化反思于一体的多维协同治理生态系统。该系统的核心目标并非追求彻底消除幻觉，因为这既无必要，也违背了大语言模型基于概率生成的本质机制，取而代之的应该是致力于增强社会对 AI 幻觉的识别、质疑与应对能力，从而阻止其伪装成共识性真理被广泛接纳。

在技术维度上，治理应实现从“杜绝幻觉”向“管理幻觉”的范式转型。其关键在于打破模型“黑箱”和增强生成过程的透明性与可验证性，具体路径包括嵌入证据关联与溯源机制，为模型输出提供可核查的信源依据；开发不确定性表征工具，使系统能够主动标识其生成内容的置信区间与局限性；强化可解释人工智能技术应用，使模型的推理路径和决策逻辑变得可审阅、可质疑（刘泽垣等，2025）。对此，Nananukul & Kejriwal（2024）提出的幻觉本体框架（HALO），即通过构建包含元数据溯源、幻觉分类及形式化查询能力的结构化模型，可实现生成过程的透明化与幻觉管理的系统化。

文化维度的治理须直面公众认知习惯与技术信任机制之间的深层互动，尤其须应对多元主体在技术使用中形成的认知惯性与价值偏好。AI 幻觉之所以能被轻易接受，既源于对技术权威的非反思性信任，也与信息过载环境下批判性文化缺失密切相关。治理的关键在于推动从“接受型认知”向“审辨型认知”的转变：加强全民 AI 素养与算法批判能力教育，使用户能够辨识技术生成内容的潜在偏差；倡导跨文化视角与多元知识观的传播，防止技术再生产单一文化叙事；鼓励公众参与知识共建与验证过程，塑造开放、包容、富有韧性的认知生态。此外，还应推动人文社会科学学者深度参与技术伦理审查和模型设计，从源头嵌入文化敏感性与价值多样性；建立社会文化影响定期评估机制，及时识别和干预其可能引发的认知偏见和文化侵蚀。

制度维度的治理重点在于构建权责清晰、多元共治的规制体系，以协调不同主体在技术实践中的利益与价值冲突。需建立独立的 AI 输出公共审计机制，系统评估幻觉的类型、频率及其社会影响；在高风险领域设置人工复核与交叉验证的强制性程序；明确开发者、部署者、用户与监管机构之间的责任分配，尤其须建立基于损害原则的责任追溯制度。同时，应推动形成社会共识导向的知识认证标准与伦理准则，使“真实性”与“合法性”的判断不再完全由技术系统决定，而是回归公共领域，通过民主商谈达成集体确认。此外，须将算法影响评估和伦理审查纳入立法程序，要求高风险 AI 系统在部署前接受跨学科专家委员会对其认知风险与社会文化影响的系统评估，从而在创新激励与认知安全之间建立动态平衡。

对生成式 AI 幻觉的治理是一场认知秩序的重建工程，它要求我们超越局部纠错和单一归责的传统治理逻辑，转而通过在技术、制度与文化三者之间建立持续互构的动态平衡，共同构筑适应人机共生时代的认知治理生态。治理的最终目标不是追求无幻觉的完美系统，而是建立一个即便在幻觉出现时仍能维持认知理性、社会信任与文化多样性的韧性体系。唯有通过多维协同治理，人类才能在与技术深度互动过程中，真正实现认知自主性的守护与文明意义的延续。

五、文化自觉语境下的人机共生

治理生成式人工智能幻觉问题，并非依靠单一维度所能应对，而需依赖技术、制度与文化三者协同的框架。技术治理贯穿“预防 – 识别 – 纠正”的全过程，既要通过模型优化降低幻觉发生概率，也要借助证据标注与不确定性提示提升可辨识性，并引入人工复核以抑制其扩散。制度治理则需构建“责任归属 – 治理方法 – 效果评估”的完整链条，明确开发者、平台与用户等各方责任，建立有效机制，并通过审计与监督确保治理实效。而在技术与制度之外，文化治理强调对认知主体性的深层关怀，致力于在幻觉无法根除的前提下帮助人类在与 AI 共生中保持价值判断与认知自主，从而维护人的主体地位。文化自觉因此成为应对 AI 幻觉的深层精神资源，也为真正意义上的人机共生奠定了价值基础。

（一）从“防止错误”到“重建自觉”

人类在面对技术带来的不确定性时，容易倾向于采取“防止错误”这一直觉性策略。传统治理始终致力于将错误发生的概率降到最低，无论是在工业时代防范机械故障与安全事故，还是在数字时代应对系统偏差与数据污染，其都试图通过技术优化与制度约束来逼近‘零差错’的理想状态。然而，生成式人工智能呈现的“幻觉”现象，迫使我们重新审视这一延续多年的治理逻辑。它揭示出一个前所未有的认知挑战：即便将表面意义上的“错误”发生频率降至极低，甚至趋近于零，也并不能真正消解认知误导的潜在风险。问题的核心已不再是统计层面的错误频次，而在于 AI 幻觉所具有的认知“欺骗性”。它往往并不以明显谬误的姿态呈现，而是隐匿于合理性表象之下，甚至通过模仿现有知识体系的术语、范式与价值倾向，获得某种表面上的合法性。此类幻觉不再是能被轻易排除的“例外”，而可能渗透成为塑造共识、影响决策乃至重构现实的一种隐蔽力量。在此背景下，若 AI 幻觉治理仍仅仅关注“减少错误”，实际上尚未真正触及认知治理的核心。因为当前的治理模式既缺乏对知识合法性生成机制的批判性考察，也未能回应技术环境下人类认知主体性所遭遇的侵蚀。此类治理模式，从根本上看，依然受制于一种停留在表面纠错的惯性思维。

因此，当代文化自觉的核心使命应当积极推动一种更具反思性与能动性的“重建自觉”。它要求我们清醒认识到：在技术持续重构认知环境的过程中，幻觉并非偶然的偏差，而是技术逻辑衍生的必然产物；其真正危险不在于数量的多寡，而在于人们是否能够保持识别、质疑与批判的能力。从这个意义上讲，文化自觉呼吁实现认知态度的根本转向：不再将 AI 幻觉简单视为亟待清除的干扰，而是把它当作反思认知建构的重要契机。通过剖析 AI 幻觉的形成机制，人类得以反观知识如何被技术媒介塑造、被权力结构渗透、被社会共识固化。进而我们应当追问：我们因何认为某些内容为“真”？知识的生产、验证与接受过程，又在何种条件下被悄然自然化？这一转向标志着治理理念的关键演进：从消极的“防止错误”迈向积极的“重建认知主体性”。它不再满足于对外部信息进行筛选与更正，而是致力于唤醒主体在认知过程中的自觉与审辨力，使人在技术编织的意义网络中仍能保持清醒、自主与能动。唯有通过这样一场深刻的文化自觉，人类才能在 AI 幻觉与真实世界交织的语境中，重建属于“人”的认知权威。

（二）人机共生中的主体性再确立

生成式人工智能的广泛应用非但未使人退居认知的边缘，反而促使人类重新审视自身在技术架构中的位置，更积极地于人机协作过程中重新确立作为认知主体的根本价值。人机共生并非意味着将判断与解释的权力让渡于算法，而是要求人类以更清醒的自觉、更主动的批判意识和更深刻的伦理责任感承担起认知治理的主导角色。在这一新型关系中，人既不是孤立的理性主体，亦非技术的附庸，而是通过与人工智能系统的持续对话、反思与校准，不断重新塑造自身的判断力和价值立场（胡泳和王昱昊，2025）。这意味着生成式人工智能所塑造的认知环境要求人类超越监督者或干预者的传统角色，转而成为阐释者、质疑者与意义的共建者。我们不仅需要理解大语言模型输出的内容，更须反思其生成路径、训练数据中隐含的偏见以及知识表征的内在局限。人的不可替代性正体现于跳出技术自洽的闭环，提出那些系统自身不会提出的诘问。例如，知识如何被特定方式组织，哪些叙事被强化、哪些现实被淡化以及背后的权力与文化动机为何。这一反思性实践正是人机认知共同体中“人”这一侧最核心的能力。

人机共生关系中主体性的重新确立，根本上有赖于人类将伦理判断与价值理性重新嵌入技术流程。生成式人工智能虽可生成内容，却无法真正对后果担责；它能模仿逻辑形式，却难以理解真理与正义的多元内涵。因此，人类必须承担起认知与伦理的双重评估使命：不仅要判断信息的准确性，更要考量其所带来的社会效应、道德正当性与情感共鸣。这一使命要求我们超越单纯的事实性核对，深入审视技术输出背后的价值预设与文化立场，进而追问其是否与人类共同的福祉和尊严相一致。与此同时，我们还需在技术设计中主动嵌入伦理约束与价值引导机制，使工具理性与价值理性在系统运作中得以协同。唯有如此，技术才能成为扩展人类认知的媒介，而不是替代人类做价值抉择的权威。

人机共生的愿景并非打造无缝衔接的技术依赖系统，而是构建一种能够持续相互审问、共同成长的认知伙伴关系。在这一关系中，人类借技术反观自身，又以自我觉察引导技术发展，从而塑造一种在人工智能时代仍以“人”为尺度的认知文明。主体性的重获正是在这一反复对话与校准中得以实现。它不是向传统自主性的简单回归，而是在与技术的深度互动中重新建立一种更具反思意识与责任伦理的人类认知身份。

（三）正面契机：从危机到转化

随着生成式人工智能逐渐应用于更多实际场景，将 AI 幻觉简单等同于“错误”的认知正在发生转变（Dumit & Roepstorff，2025）。目前业界正形成一种更具区分度的新认知：尽管大语言模型的幻觉源自其概率生成机制，但在创意写作、策略推演或头脑风暴等场景中可以提供意想不到的关联与启发，扮演着“创造性偏离”的积极角色（殷杰，2024）；而在医疗、司法、事实核查等对准确性要求极高的领域，需要通过检索增强生成、智能体协作、不确定性量化等技术手段，对幻觉带来的实质性风险进行约束和管理。

除了技术层面的应对，AI 幻觉问题实际上反映出人类认知与机器信息处理逻辑之间的深层差异。它迫使我们重新思考一些根本问题，让我们看到人类自身知识体系中那些未经审视的前提。因此，我们或许应该跳出完全消除或分区管控这种非此即彼的思维。AI 幻觉可以被视为推

动我们认知体系演进的一种内在动力。它提醒我们，真正的进步不在于追求一个永远正确的人工智能，而在于建立一种能够接纳不确定性，并持续学习和自我修正的认知方式。这不仅需要技术上的改进，更呼唤人类认知范式的深层转变：从追求绝对确定的真理观转向拥抱开放、可修正的知识生态；从被动接受技术输出转向保持清醒的批判性思考；从依赖外部权威验证重返对自身判断力的信心与培养。从这个角度看，AI 幻觉不再仅仅是一个技术问题，更是一个促使我们不断反思、走向成熟认知社会的契机。它邀请我们重新审视理性的边界、知识的本质以及人类在技术时代的主体性位置。我们或许能够通过理解 AI 幻觉，最终在人与技术的协作中找到更稳健的平衡点，构建出一种既拥抱技术潜力又不丧失人文智慧的共生方式。

六、结论与展望

生成式人工智能的幻觉问题，远不止是一项技术层面的缺陷，更是深刻嵌入当代知识生产与认知结构之中的系统性现象。它既源自大语言模型内在的概率性生成机制，也成形于特定文化语境中被不断强化的信任惯习与认知偏好。本文试图揭示：AI 幻觉本质上构成一种“认知僭越”，不仅干扰个体的信息接收与判断，更在深层次上动摇人类长期以来所依赖的真实性标准与知识合法性基础。

面对这一多重属性的挑战，传统的纠错式治理已显得力有未逮。我们主张，应超越单纯的技术修复和外部管控，转向“技术 – 制度 – 文化”三维协同的治理路径。技术维度上，应推进可验证、可解释的模型设计，增强生成过程的透明度与溯源性；制度层面，须构建权责清晰、多方参与的责任框架，确立适应人机协作时代的认知伦理规范；文化向度上，则应培养公众的 AI 素养与批判意识，重塑技术环境中的认知主体性与文化自觉。这三者并非彼此割裂，而是相互支撑、动态调适的整体性治理生态。其目标不是追求无法实现的“零幻觉”系统，而是通过增强社会对 AI 幻觉的识别、反思与应对能力，构建一种能够包容不确定性、并在错误中持续学习的认知韧性。

展望未来，生成式人工智能将更深入地融入知识生产、文化传播和日常决策中，成为人类认知活动中不可忽视的“他者”。AI 幻觉或许无法被根除，但它可以从被动的风险源转化为积极的反思契机。因此，我们在推进技术迭代的同时更应开展一场深层的认知与文化转型：从追求确定性的传统知识观转向接纳开放、可修正的认知生态；从依赖外部权威重返基于批判性思维和伦理判断的主体性位置；从被动应对技术冲击转向主动参与认知秩序的共建。

最终，人类能否在人工智能时代维护认知的自主与尊严，并不取决于是否能够消灭 AI 幻觉，而在于我们是否能够以清醒的自觉、持续的审辨和深厚的文化责任感与技术共同进化。人机共生的真正实现需要的不仅是更聪明的算法，也需要更智慧的人类。

参考文献

崔星璐、姚长青，2025：《我国人工智能系统分级透明制度的构建路径：基于欧盟 < 人工智能法案 > 风险评

估模式的研究》,《情报杂志》第 8 期。

胡泳、王昱昊，2025 :《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》,《南京社会科学》第 3 期。

黄立赫，2025 :《生成式人工智能中“信息茧房”的生成与应对》,《图书情报工作》第 8 期。

刘明君，2025 :《信息过载时代内容筛选的人机协同模式及其优化》,《出版广角》第 7 期。

刘永芳、许科、尚雪松，2025 :《理性观视野中的人类合作行为》,《心理科学》第 2 期。

刘永谋、孙瑞璇，2025 :《非主体 : 人工智能的定位与治理》,《新经济》第 6 期。

刘泽垣、王鹏江、宋晓斌、张欣、江奔奔, 2025 :《大语言模型的幻觉问题研究综述》,《软件学报》第3期。

[法] 米歇尔 · 福柯，舒炜、王晨晨译，2019 :《规训与惩罚》，生活 · 读书 · 新知三联书店。

[德] 马丁 · 海德格尔，孙周兴译，2020 :《演讲与论文集》，商务印书馆。

[美] 尼尔 · 波斯曼，何道宽译，2007 :《技术垄断 : 文化向技术投降》，北京大学出版社。

王鸿宇、蓝江，2025 :《智能时代知识生产及其结构性风险——基于 DeepSeek 生成式人工智能技术的哲学反思》,《理论与改革》第 3 期。

许可，2025 :《人工智能法律规制的第三条道路》,《法律科学》(西北政法大学学报) 第 1 期。

殷杰，2024 :《生成式人工智能的主体性问题》,《中国社会科学》第 8 期。

赵越、师小雅、苏广玲，2024 :《大语言模型技术下人工智能的偏见与规制——以 ChatGPT 为例》,《新媒体与社会》第 4 期。

周妍、沈天健，2024 :《生成式人工智能视域下虚假信息的层级化运作机理与治理》,《编辑之友》第 8 期。

Cheng, A., Nagesh, V., Eller, S., et al.，2025，“Exploring AI Hallucinations of ChatGPT: Reference Accuracy and Citation Relevance of ChatGPT Models and Training Conditions”. *Simulation in Healthcare : Journal of the Society for Simulation in Healthcare*, 10. 1097/SIH.0000000000000877.

Dumit，J., and Roepstorff，A., 2025，“AI Hallucinations are A Feature of LLM Design, Not A Bug”，*Nature*, 639(8053), 38.

Nananukul, N., and Kejriwal, M., 2024, “HALO: An Ontology for Representing and Categorizing Hallucinations in Large Language Models”. *Disruptive Technologies in Information Sciences*, 13058, 86–100.

Pak, R., Rovira, E., and McLaughlin, A.，2024, “Polite AI Mitigates User Susceptibility to AI Hallucinations”，*Ergonomics*, 68(10), 1735–1745.

Vaswani, A., Shazeer N, Parmar N, et al., 2017, “Attention Is All You Need”, *31st Conference on Neural Information Processing Systems* (NIPS 2017), 1–11.

Technology and Culture: Cognitive Transgression in Generative AI Hallucinations and Governance Transformation

CAO Yongping

Abstract: Hallucinations in generative artificial intelligence are not simply technical flaws but manifestations of cognitive transgression that reshape the conditions of knowledge production and governance. They not only challenge human verification mechanisms and the credibility of knowledge systems, but also erode social trust. Existing governance responses have largely emphasized technical fixes and regulatory measures, while paying insufficient attention to the cultural dynamics that enable the generation and diffusion of hallucinations. This paper adopts a dual lens of philosophy and culture to frame hallucinations as outcomes of the interaction between technological logics and cultural narratives. On this basis, it proposes the concept of cognitive transgression to explain how generative AI hallucinations gain legitimacy through technological authority, cognitive inertia in contexts of information overload, and narrative biases. The analysis highlights the need to move beyond error correction toward a governance paradigm that integrates technical, cultural, and institutional dimensions: enhancing transparency and verifiability, promoting critical literacy and algorithmic literacy, and developing accountable, pluralistic mechanisms of co-governance. The paper argues that the ultimate goal of governance is not to eliminate hallucinations entirely, but to reconfigure cognitive order, so that hallucinations can be transformed into drivers of human-AI symbiosis and renewed epistemic resilience.

Keywords: Generative Artificial Intelligence; Hallucination; Cognitive Transgression; Cultural Reflexivity; Governance Transformation

【责任编辑：王韵清】