

# 行动者网络理论视角下 AI 知识幻觉的生成机制与协同治理路径研究

徐延民<sup>1</sup> 胡晓萌<sup>2</sup>

(广东海洋大学马克思主义学院, 湛江 524088)<sup>1</sup>

(广州大学马克思主义学院, 广州 510006)<sup>2</sup>

**[摘要]** 生成式人工智能的勃兴在重塑科学传播范式的同时, 其衍生的“知识幻觉”现象正在深度侵蚀科学传播的公信力基础。本研究基于拉图尔行动者网络理论分析框架, 将 AI 知识幻觉视为一种在异质性行动者网络中由于关键行动者的目标偏离与利益冲突, 导致“转译”过程中发生系统性失真或“背叛”的产物。这根植于四重结构性失调: 源头网络中数据“铭刻”的历时性偏见与算法目标异化导致知识源污染; 纠错网络内反馈机制的制度性缺位与时间异步性矛盾阻碍错误修正; 转译链条面临专业话语降维损耗与异质知识网络的技术暴力性交叉污染; 责任网络因技术“黑箱”遮蔽与权责离散陷入治理真空。基于此, 本研究提出协同治理的四维重构路径: “源头净化”通过多中心知识认证体系与区块链溯源技术筑牢数据根基; “过程疏通”借力智能化监测系统与动态知识库构建闭环纠错网络; “转译优化”采用语境感知算法与人机双重校验机制保障知识保真度; “责任锚定”依托法律赋权与透明化算法披露厘清多元主体责任边界。以期加深对 AI 知识幻觉生成机制的社会—技术本质的理解, 进一步推动人机协同、稳健可信的“AI+ 科普”生态建设。

**[关键词]** 行动者网络理论 人工智能 知识幻觉

**[中图分类号]** TP18; N4 **[文献标识码]** A **[DOI]** 10.19293/j.cnki.1673-8357.2025.05.003

伴随生成式人工智能 (Generative Artificial Intelligence, 生成式 AI) 大语言模型 (Large Language Model, 以下简称大模型) 的发展, DeepSeek、ChatGPT 等生成式 AI 大模型 (以下简称 AI 大模型) 在知识普及、科学传播上潜力巨大、大有可为, 但也逐渐暴露出其知识幻觉的问题, 即其提供的知识表面上看起

来合理, 但实际上却并非正确, 使得 AI 大模型在知识密集型应用场景下的可靠性和有效性问题备受质疑<sup>[1]</sup>。知识幻觉会直接带来 AI 大模型在知识密集型应用场景下 (包括知识科普、医疗保健、辅导教育等) 的可靠性质疑。尤其是科普领域, 若 AI 频频“说错话”, 可能让谣言和误解乘虚而入, 进一步加剧科

收稿日期: 2025-06-21

基金项目: 国家社会科学基金 (24FYB017); 广东海洋大学科研启动费资助项目 (R20074); 广东省高等教育学会 2025 年高等教育研究课题 (25GYB043)。

作者简介: 徐延民, 广东海洋大学马克思主义学院讲师, 研究方向: 技术哲学、技术社会学, E-mail: xym413@yeah.net。胡晓萌为通讯作者, E-mail: huxiaomengnan@foxmail.com。

学与公众之间的认知鸿沟，甚至可能引发人们对新科技的不信任和抵触，阻碍 AI 技术在社会生活中的应用。正如习近平总书记在中共中央政治局第二十次集体学习时强调：“面对新一代人工智能技术快速演进的新形势，要充分发挥新型举国体制优势，坚持自立自强，突出应用导向，推动我国人工智能朝着有益、安全、公平方向健康有序发展。”<sup>[2]</sup>因此，确保人工智能发展安全、可控、可靠至关重要，这是共同推动人工智能造福人类社会的前提。

## 1 理论基础与分析框架

当前学者对 AI 知识幻觉的研究大多集中于技术维度<sup>[3]</sup>，如训练数据偏差、模型结构的局限性等<sup>[4]</sup>，却忽视了知识幻觉的社会维度，对 AI 知识幻觉现象的评估和改善优化存在明显不足<sup>[5]</sup>。AI 知识幻觉现象不仅是单纯的技术缺陷，更是人与技术、数据与算法、用户与系统之间复杂互动的产物。因此，仅从技术层面“头痛医头、脚痛医脚”是不够的，本研究超越技术决定论，以行动者网络理论为分析框架，将 AI 知识幻觉置于一个由人类行动者与非人类行动者共同构成的复杂社会—技术网络中进行考察，以期为技术开发商提供改进 AI 大模型的方式，为科普工作者、政府、公众参与 AI 治理提供思路，通过多方联动，将 AI 从“误导者”变为助力科普的“真帮手”，促进公众认知提升。

### 1.1 AI 大模型知识幻觉现象研究

知识幻觉（Knowledge Hallucinations），即 AI 大模型生成的内容虽表面流畅但与事实不符，甚至包含逻辑错误或概念混淆的现象，已成为影响人工智能系统可靠性的核心问题之一<sup>[6]</sup>。这种现象就像一个学生在考试中“胡编乱造”，答案看似完整，却漏洞百出。知识幻觉大致可分为两类，即事实性幻觉和忠实

性幻觉。其中，事实性幻觉表现为事实捏造，与事实不一致；忠实性幻觉表现为指令性错误、逻辑性错误，同上下文不一致<sup>[7]</sup>。在科普实践中，这种问题尤其棘手，因为公众往往缺乏专业背景，难以辨别 AI 输出信息的真伪，一旦被误导，纠正成本极高。AI 知识幻觉是不可避免的事实，降低幻觉率并提升 AI 大模型的可靠性和普适性是未来的重点方向<sup>[8]</sup>。

### 1.2 行动者网络理论与 AI 知识网络的 ANT 模型拓展

行动者网络理论（Actor–Network Theory, ANT）由法国学者拉图尔（Bruno Latour）等人于 20 世纪 80 年代提出，强调社会结构网络并非单纯由人类行为决定，而是由人类（如开发者、用户）和非人类（技术物、文本、科学概念、组织机构等）共同构成的异质网络<sup>[9]</sup>。在这些网络中，任何行动者的身份、能力和意义都不是预先固定的，而是在与其他行动者发生关联、互动的过程中被“转译”和“铭刻”出来的<sup>[10]</sup>。通过这一理论视角，可以将 AI 知识幻觉看作是网络中各“角色”在互动过程中由于目标偏差、信息失真、责任模糊而形成的综合产物。

AI 知识生产传播网络系统是异质性组织之间的行动耦合过程，涉及权力分配、价值观念等<sup>[11]</sup>。本研究将 AI 知识生产置于社会—技术网络中进行系统考察，这一框架突破了技术决定论的局限，强调行动者之间的动态关系而非其固有特性，从而从“幕后推手”的互动中挖掘产生幻觉的根源。

其结构分析框架由四大场域构成（见表 1）。一是算力—数据场域，由 AI 运行需要的算力设施、数据管道、硬件部署等非人行动单元构成，是 AI 知识生产的“地基”，它们决定模型质量。一旦数据管道中掺入了劣质数据，AI 所输出信息的准确性将大打折扣。二是算法—模型场域，由模型架构、奖励模

型等异质行动者构成。算法成为“铭刻”了开发者意图和价值的载体<sup>[12]</sup>。比如奖励模型可能优先优化用户满意度而非事实准确性，导致 AI 倾向于生成“讨好型”而非“真实型”内容。三是人一机交互场域，由接口设计、提示工程、反馈循环等人类与非人类行动者之间的互动构成。用户在与 AI 系统互动过程中，难免因为提示设计不合理导致生成偏见内容，类似“导游和游客”之间互动，倘若双方沟通不畅将会导致“误会”。四是社会一制度场域，这包括评价标准、伦理规范、监管协议等。AI 知识生产的正当性及可接受性，正是通过这些规范制度等措施来保障的。缺乏统一的事实核查标准以及伦理规范将会导致我们忽视知识幻觉。

AI 知识生成网络是非中心化、非线性系统，强调异质行动单元的平等能动性。在 AI 知识生成网络中，这种平等性并非指权力分配的绝对均衡，而是强调人类行动者（开发者、用户、监管者）与非人类行动者（数据、算法等）均具备独立的能动性与转译能力，是通过相互征召、协商与互动共同塑造出来的结果。拉图尔强调行动者的能动性并非取决于其物质属性，而是体现在其“转译”

其他行动者目标与利益的能力上。在 ANT 视域下，“转译”是构建和稳定网络的核心动力，它通过一系列协商、说服与联盟，将异质行动者的利益绑定在一起，形成协同。然而，“转译”并非一个必然成功的过程。当行动者的目标无法对齐、利益发生冲突或信息在传递中被扭曲时，就会发生“转译的失败”或“背叛”。此时，行动者仍在进行转译，但其行为却在削弱而非巩固网络最初设定的目标。AI 知识幻觉正是在此意义上，被视为数据、算法、用户等多元行动者在一系列不成功的“转译”中累积而成的系统性结果，它体现了网络行动者对“传播准确知识”这一集体目标的“集体背叛”。比如数据偏见在“征召”阶段被嵌入，交互中的用户误解在“动员”阶段扩散，制度规范在“利益赋予”阶段可能滞后。AI 知识幻觉并非单一环节的失误，而是数据、算法等多重行动者在转译过程中目标偏差、信息失真的综合结果。从算力—数据场域的“铭刻”和“征召”，算法—模型场域的“输出转译”，到人—机交互场域的“问题呈现”，再到社会—制度场域的“动员”和“规范”，AI 知识幻觉成为整体网络转译过程的集体产物。

表 1 四大场域与行动者框架

场域	人类行动者	非人类行动者	ANT 角色
算力—数据	数据标注员、云计算工程师等	训练数据集、GPU 集群等	征召者（将原始数据转化为可用资源）
算法—模型	算法设计师、调参工程师等	神经网络、算法模型等	转译者（将数据模式转化为预测输出）
人一机交互	UX 设计师、终端用户	AI 交互系统、反馈日志系统等	动员者（维持用户与 AI 的互动循环）
社会—制度	AI 法律起草者、监管机构等	法律法规、行业标准等	强制通行点（定义网络合法性边界）

我们在运用行动者网络理论时，并非机械地坚持绝对的“对称性原则”。ANT 的对称性原则并非要抹杀人类的主观能动性，而在于强调在网络构建中，任何行动者的力量是在与其他行动者关联和协商中动态生成的。在 AI 知识幻觉的生成网络中，数据、算法等非人类行动者通过其独特的能动性与人类行动者塑造了网络体系。然而，我们在探讨

治理网络时面对现实的非对称性，即责任的归属与伦理的裁决在当前的社会—法律框架下，仍然是人类行动者无法让渡的核心角色。因此，这需要在“诊断问题”时遵循对称性，以充分揭示非人类行动者的能动性；在“提出对策”时则聚焦于人类行动者如何构建新的网络规则、引导和约束其他行动者，从而构建权责相对清晰、人机协同更为稳健的新

网络。这种从“对称分析”到“非对称治理”的转换，本身就是对行动者网络理论在特定治理情境下应用的批判性反思。

## 2 AI 知识幻觉的网络生成机制

作为一种复杂的现象，AI 知识幻觉的生成并非单纯的技术缺陷，而是嵌入社会—技术网络中的多重行动者互动结果。基于 ANT 视角，我们可以将 AI 知识幻觉的产生过程，看作是网络中各个“角色”（行动者）在目标传递与互动过程中，出现偏差的结果。其主要生成机制可归纳为以下 4 个方面（见图 1）。

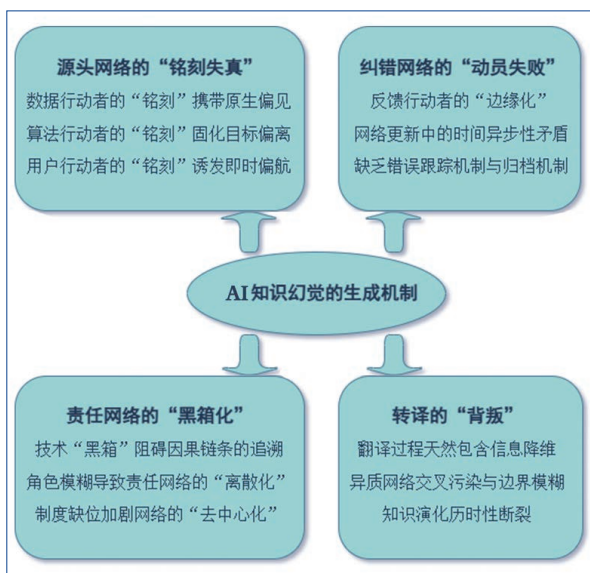


图 1 AI 知识幻觉的生成机制

### 2.1 源头网络的“铭刻失真”：行动者目标偏离与数据污染

在 AI 知识生成的网络中，核心异质行动者（如数据、算法、用户）的本意是传递和生成知识，但其内在局限或目标偏差，在“转译”过程中造成了系统性“铭刻失真”。这种失真是行动者网络复杂交互的必然结果，这为幻觉的产生埋下伏笔<sup>[13]</sup>。以下从 3 个具体维度拆解这一问题。

其一，数据行动者的“铭刻”携带原生偏见。在 ANT 视角下，训练数据并非被动“喂料”，而是携带着特定历史、语境与价值

的能动性“代言人”。作为历史与社会偏见的“代言人”，成功地“说服”算法模型接纳其携带的错误叙事，从而在网络源头成为滋生幻觉的温床。当 AI 网络大规模“征召”来自互联网的数据行动者时，这些数据所“铭刻”的瑕疵——如过时的科学论断、未经验证的民间叙事，乃至伪科学信息——被无差别地纳入模型。以气候变化议题为例，若网络“征召”了大量否认人为因素的语料，那么“伪科学”便被成功“铭刻”为 AI 知识结构的一部分，导致其在后续互动中生成与“真科学”相悖的结论。这种源于数据层面的“铭刻失真”具有高度隐蔽性，因为互联网这一“行动者储备库”本身异质混杂且缺乏质量控制，使得即便经过清洗，风险也难以根除，直接损害了 AI 科普网络的初始可信度。

其二，算法行动者的“铭刻”固化了目标偏离。算法作为开发者意图的“铭刻”载体，其核心目标并非追求绝对的“真实”，而是生成统计上最“貌似合理”的文本序列。这种被“铭刻”的“模仿”而非“理解”机制，使算法成为高效的“概率文书”，而非“逻辑学者”。当其面对训练数据中未被充分覆盖的复杂或前沿科学问题时，其内在的统计关联逻辑便会取代事实因果逻辑，通过“自由组合”或“生搬硬套”的方式“创造”出看似流畅但实则谬误的内容。例如，将量子力学的不同概念进行不当关联，正是算法行动者忠实执行其“追求文本连贯性”这一“铭刻”指令所导致的必然结果，这种目标偏离是幻觉生成的关键机制。

其三，用户行动者的“铭刻”诱发了即时偏航。在人机交互这一微观网络中，用户及其输入的“提示词”成为具有强大引导力的中心行动者<sup>[14]</sup>。当用户提出一个本身带有预设、歧义或错误前提的问题时，该提示词就如同一个“剧本”，将 AI 行动者“征召”

进一个特定的、偏离事实的叙事轨道。AI 为完成其“迎合用户意图”这一被“铭刻”的次级目标，便会生成符合用户错误预期的幻觉内容<sup>[15]</sup>。同时，AI 系统被开发者“铭刻”上的过度自信的语言风格，又进一步强化了幻觉的可信度，使得缺乏科学素养的用户难以识别谬误。因此，幻觉成为用户与 AI 在即时互动中共同“协商”和“构建”的产物，显著加速了不实信息的网络扩散。

## 2.2 纠错网络的“动员失败”：反馈行动者的缺位与网络阻滞

在 ANT 框架下，知识生产系统本质是由人类行动者、技术行动者、制度行动者等异质行动者构成的复杂网络，其中纠错机制作为维护网络健康的关键环节，本应发挥“动员联盟”功能，然而当前 AI 知识网络中的纠错行动者被严重边缘化，导致网络自我修复功能受阻。这种“动员失败”导致 AI 知识幻觉长期存在并难以消解。以下从 3 个维度剖析这一“动员失败”机制。

其一，反馈行动者的系统性“边缘化”。在 ANT 视域下，用户作为关键的人类行动者，本应在纠错网络中扮演“问题识别者”与“信息转译者”的角色，但现实中，这些行动者常被排除在网络运作的核心之外。大多数 AI 系统要么完全缺乏有效的反馈通道，要么设置了形式化的反馈机制但未将其真正纳入知识生产网络的循环。这种反馈行动者的“结构性缺位”，不仅是技术设计的失误，更是一种网络权力不平衡分配的表现。AI 系统开发者作为强势行动者，通过控制网络边界和互动规则，有意或无意地弱化了用户在纠错网络中的发言权与影响力，最终导致幻觉问题在缺乏有效制衡的封闭网络中不断累积。

其二，异质行动者在网络更新中的时间异步性矛盾。在行动者网络中，模型更新与信息传播这两类过程由不同的时间逻辑所主

导，形成了明显的“异步性冲突”。模型更新作为一个复杂的多行动者协同过程，往往需要数周甚至数月的漫长转译链条，而互联网信息传播则遵循“即时扩散”的网络逻辑。这两种时间行动者之间的张力，导致错误信息可能在模型完成自我修正之前，已经通过社交媒体等节点被成千上万的用户行动者接收并内化，形成稳定的认知网络。随着暴露时间的延长，错误信息在社会网络中的根植性增强，纠错的网络成本呈指数级上升，形成典型的“路径依赖”困境。这种行动者的时间异步性，削弱纠错网络的动员效能。

其三，缺乏系统的错误跟踪机制与归档机制。AI 系统的理想状态应建立纠错数据库，应当记录已发现的知识幻觉问题并积累到纠错数据库中，成为 AI 系统的学习资源。但目前大多 AI 系统缺少纠错系统跟踪机制，导致同样类型的知识幻觉反复出现，系统不能从过去的错误中学习。AI 系统的这种“健忘”导致系统纠错陷入无效“劳作”。与此同时，纠错反馈机制的缺失可能削弱用户的参与积极性。用户如果多次反馈纠错未起作用，就会产生“反馈无用”心理进而降低参与纠错的积极性。换言之，缺乏纠错机制的 AI 系统，就像缺乏免疫系统的有机体，可能当错误积累到一定程度后被人们所抛弃。

## 2.3 转译的“背叛”：知识在链条传递中的意义漂移与网络失序

基于 ANT 框架，AI 知识幻觉的生成可被视为在复杂知识网络构建过程中，由于多重“翻译”环节的失真所引致的现象。在 ANT 视域下，转译一般分为问题化、利益赋予、动员和招募 4 个阶段。“转译”不仅指语言转换，更指涉知识、利益与权力在异质行动者（包括人类与非人类）之间传递、转化与重新配置的过程<sup>[16]</sup>。在 AI 驱动的知识生产与传播网络中，知识从原始文本语料到计算机可处理

的数字表征，再到模型内部的复杂运算与模式生成，最终通过人机交互界面输出为用户可感知的文本或多模态内容，这一系列过程均可被视为连续的“翻译”链条。在此链条的任一节点，若行动者（如数据标注者、算法设计者、模型本身、用户）的目标、认知框架或技术约束未能有效对齐，或在“翻译”过程中发生信息损耗、语义漂移或意义的非预期重构，均可能导致知识的失真，进而表现为“知识幻觉”。

其一，从专业知识到通俗表达的翻译过程天然包含信息降维。科学知识在其固有的专业形态下，往往具有高度的复杂性、精确性与语境依赖性。将其转化为公众易于理解和接受的通俗表述，本身即构成一种极具挑战性的“降维翻译”。即使对于经验丰富的人类科普创作者而言，在确保科学性的前提下实现有效传播亦非易事。相较之下，当前 AI 系统在执行此类知识层级转换任务时，由于其在深层语义理解、概念抽象及语境适应性等方面的固有局限，更容易在“降维翻译”这一不忠实的转译策略中，造成信息的过度简化、核心内涵的偏离或关键细节的遗失。同时，失真还表现在 AI 系统内的多层次表征转译。AI 系统采用多层神经网络，知识信息在传递过程中经过多层转译。AI 系统内的“多级转译”，导致最终输出与初始知识信息的关系复杂化。即使语义的表征在最先进的语言模型中，也会在传递过程中出现“退化”现象。

其二，异质网络的交叉污染与边界模糊。知识网络的边界流动性是创新的源泉，而 AI 系统的强制融合却将这种流动性异化为技术暴力性的交叉污染。真正的知识系统是动态平衡体系，既尊重异质网络的自然流动，又警惕技术暴力对边界的异化。ANT 强调网络的异质性与边界的流动性，在 AI 系统中，不

同知识领域作为各自独立的行动者网络，在模型训练过程中被迫进行了非自然的交叉与融合。这些交叉污染不仅仅是简单的事实错误，更是异质知识网络在算法空间中不当“翻译”的产物，反映了 AI 系统在处理复杂知识网络边界时的固有局限，以及当前数据训练范式对知识网络完整性与边界保护的忽视。对于 AI 系统而言，关键不在于追求“绝对清晰的边界”，而在于建立“有语境感知的边界协商能力”，使 AI 系统真正成为知识网络中负责任的“异质行动者”，而非破坏性的“外部入侵者”。

其三，知识演化的历时性断裂问题在 AI 知识生产中较为突出。从 ANT 视角出发，科学知识并非静态，而是处于不断动态建构中的存在。AI 驱动知识生产方式发生改变，推动知识的快速迭代与更新<sup>[17]</sup>。然而，当前 AI 系统在知识处理上存在“时间扁平化”的现象，其生成的知识内容往往混合了不同时期的科学认知，既包含当前主流观点，又掺杂着已被修正的旧有理论。这不仅造成公众认知的混乱，更削弱了科学传播应有的历史纵深感。要解决这一问题，需要在数据标注环节引入时空维度，建立知识演变的时空坐标系，助力 AI 系统能够识别并呈现科学认知的历时性特征。这不仅有助于提升 AI 生成内容的准确性与可信度，也有助于推动公众对科学知识的理解更全面。

#### 2.4 责任网络的“黑箱化”：行动者角色的模糊与权责离散

在 AI 驱动的知识生产体系中，责任归属问题日益呈现出“黑箱化”特征。这种现象不仅源于技术本身的复杂性，更是多重行动者网络交互过程中权责关系失序的结果显现。其核心困境主要体现在以下维度。

其一，技术“黑箱”阻碍因果链条的追

溯。AI 大模型，尤其是深度学习模型，其内部运作机制高度复杂且非线性，构成了典型的“技术黑箱”。当系统产生“幻觉”或错误输出时，即便开发者也难以清晰追溯其具体生成路径。这种不可解释性使得责任认定陷入“源头难寻”的困局，如同医生无法确诊病因则难以有效施治。一个无法被观察和理解的技术网络，自然难以有效修复和明确问责，导致许多治理措施易停留于表层，难以触及根本。

其二，角色模糊导致了责任网络的“离散化”。AI 知识网络由数据供应方、算法开发者、平台运营方、用户等多元行动者共同构成。理想状态下，他们应形成一个权责清晰、紧密耦合的责任网络。然而，“黑箱”的存在为行动者之间的责任推诿提供了“便利”。开发者可将责任“转嫁”给上游的数据行动者；平台方则可将责任“归因”于下游的用户行动者。每一个行动者都只固守自己狭隘的角色定义，拒绝承认其行为对网络整体后果的连带责任。这导致原本应相互连接的责任网络分解为一个个孤立、离散的“行动者孤岛”，无法形成有效的协同问责机制。

其三，制度缺位加剧了网络的“去中心化”。一个稳定的网络需要一个或多个“中心行动者”（如法律、伦理框架规范）来定义和强制执行网络规则。当前，针对 AI 知识幻觉的法律与伦理框架这一关键的制度行动者尚处于“被招募”的早期阶段，其力量薄弱，无法有效扮演网络中心的角色。法律责任的模糊、举证的困难，使得责任网络缺乏一个最终的“仲裁者”和“稳定器”，呈现出一种无序的“去中心化”状态<sup>[18]</sup>。AI 系统平台的自主决策与人类的有效监督之间形成责任真空<sup>[19]</sup>。在商业利益和公共责任冲突的背景下，开发者和平台利益相关者往往追求市场扩张

的经济目标压倒科学传播的公共责任目标，进一步加固了“黑箱”，侵蚀了整个科普网络赖以生存的社会信任。

### 3 AI 知识幻觉的协同治理：构建稳健的行动者网络

在 AI 深度介入知识生产的大环境下，AI 知识幻觉问题已演变为典型的社会技术难题。我们在行动者网络理论基础上，需要实践性地回归具体应用环节<sup>[20]</sup>。“四维协同”治理框架从源头质量保障、过程反馈矫正、知识转译优化与责任框架建构 4 个方面展开，其目的在于降低 AI 知识幻觉风险，构建稳健知识生产网络系统，具体思路如图 2 所示。

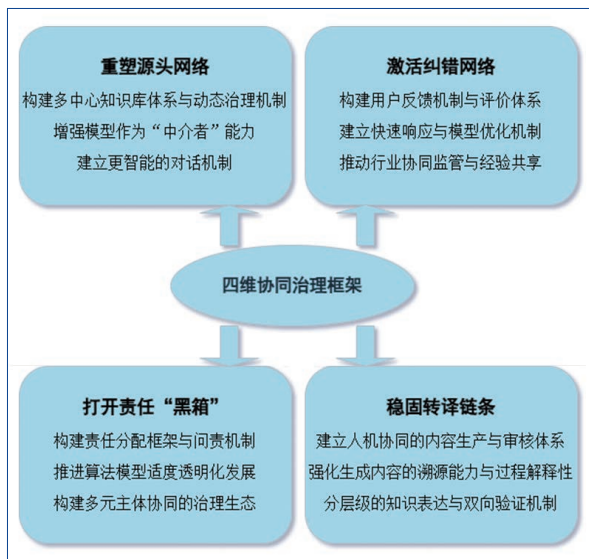


图 2 四维协同治理框架

#### 3.1 重塑源头网络：从“铭刻”环节保障行动者的目标对齐

AI 知识生成的可靠性依赖数据基础、算法架构和交互设计等要素。我们从数据、算法与交互 3 个维度进行分析。

其一，在数据层面，应构建具有“强铭刻性”的多中心知识库体系与动态治理机制。稳定的网络需要将行动者的利益与目标“铭刻”到物质性载体中。科研机构、专业学会

与科普组织等多元主体应形成“知识资源动员联盟”，通过共同认证的“铭刻”标准提供经过严格验证的结构化科普语料，为 AI 模型训练提供高可信度的知识基础。这种多中心协作将数据质量控制从单一技术问题转化为社会协作网络，形成知识生态中的“信任锚点”。与此同时，应发展数据溯源与质量评估技术体系，通过区块链等分布式技术构建数据来源与修订历史的透明化追溯机制，使知识的演变历程成为可见的“铭刻轨迹”。特别值得关注的是知识更新机制的建立，通过与权威学术数据库、科研机构的动态对接，实现知识库的时效性维护，减少因网络中“时间异步性铭刻”而产生的系统性幻觉。

其二，在算法层面，应增强模型作为“中介者”而非简单“中间物”的能力。也就是说，算法设计需要超越简单的信息传递功能，实现真正的知识转译，即实现从“信息传递者”到“知识转译者”的转变。根据拉图尔的分， “中间物”仅传递意义而不改变其内容，而“中介者”则转译并重构所传递的内容<sup>[21]</sup>。AI 大模型应从简单的中间物转变为具备认知审慎的中介者，整合知识图谱、因果推理等技术强化对概念间本体关联的把握，以减少仅基于表面相关性的错误推断。当面对信息不足或存在争议的领域时，系统应具备“元认知铭刻”能力，通过概率化表达或明确的限定语来传达知识的不确定边界，而非强行生成确定性回答。多模态融合框架为知识的可靠性提供了异质性验证路径，通过文本、图像等多元“铭刻”方式的交叉验证，构建 AI 系统的“复合铭刻体系”，避免单一模态造成的认知偏差。

其三，在人机交互层面，应建立更智能的对话机制。用户提问的质量与精确度直接影响 AI 回复的质量，可通过智能化引导工具

辅助用户准确表达需求。输出界面中明确标识“AI 生成内容”并融入交互式核查功能，允许用户对关键信息展开质询，形成“质疑—回应—验证”的闭环交互模式。这种互动模式的优化需要与用户教育相结合，提升公众对 AI 生成内容的批判性解读能力，培养用户成为知识网络中的“主动校正者”而非被动接受者，这种人机协同的知识验证网络是构建健康 AI 科普生态的基础<sup>[22]</sup>。只有当用户具备辨别能力，才能有效规避知识幻觉风险，实现人机协同的知识生产。

### 3.2 激活纠错网络：对反馈行动者的有效“动员”与激励

有效治理 AI 知识幻觉现象，亟须构建一个多方联动、动态响应的纠错体系。当前，反馈渠道的碎片化导致问题响应迟滞，修正流程的滞后性助长误导信息扩散，行业协作的不足削弱了整体纠错效能。解决上述问题的关键在于融合用户、专家智慧与技术力量，共同打造“监测—响应—优化—共享”的闭环网络。

其一，构建分层化、激励性的用户反馈机制与评价体系。传统的反馈渠道常因分散、低效而难以形成有效闭环。为此，应建立标准化、结构化的反馈系统，整合普通用户、领域专家与科普工作者等多元主体的评价渠道，构建层次化的内容质量监测网络。尤其是反馈激励机制的设计，通过积分奖励、专业认证等多元化激励手段，提升用户参与纠错的积极性。同时，结合众包与专家评审的混合验证模式，能够平衡反馈框架体系，在大众发现问题的基础上，由专业人士进行深度审核，形成科学的质量控制体系。

其二，建立幻觉内容的快速响应与模型迭代优化机制。AI 知识幻觉的发现与修正往往存在时间差，这种滞后性客观上导致误导

信息传播扩散<sup>[23]</sup>。为降低风险，应发展智能化的幻觉监测系统，结合异常检测算法与用户主动报告，构建“系统自检”与“用户反馈”相结合的双重监测网络，主动识别潜在幻觉内容。同时，系统化地建立幻觉案例知识库，对已发现的幻觉进行类型学分析，深入挖掘其认知生成机制，并将这些案例作为负向样本或校正指令反馈到模型训练循环中，实现能力精准提升。换言之，需优化“发现—验证—修正—部署”的迭代周期，通过在线学习技术与增量训练等方法，实现模型参数的动态更新，从而缩短纠错的响应时间，最大限度遏制误导信息的传播范围。

其三，推动行业协同监管与经验共享。AI 知识幻觉的治理需要行业整体的集体智慧，应鼓励不同 AI 开发机构、应用平台之间建立匿名的幻觉案例共享数据库，在保护商业秘密的前提下，促进相关经验的交流互鉴。这种协同治理机制不仅有效避免各平台重复犯错，还可以形成整体行业的“集体免疫”效应，共同筑牢科普内容的质量与可靠性的防线。

### 3.3 稳固转译链条：构建人机协同的“保真”传递

在科普传播过程中，AI 作为专业知识与大众理解间的中间变量，其核心挑战在于如何既确保科学性又兼顾可理解性。实现这一目标，需要构建人机协作的“双重校验”机制，从协同生产、内容溯源与分层表达入手，保障知识传递的“保真”度，构建工具理性和价值理性协同的科普传播范式<sup>[24]</sup>。

其一，建立人机协同的内容生产与审核体系。单纯依赖 AI 自主生成往往难以平衡专业准确性与传播亲和力。引入人类专家参与，进一步构建“AI 初稿生成—专家专业审校—AI 辅助优化”的协作流程。具体而言，AI 负责信息的初步整合与结构化草稿生成；人类专家则聚焦于科学知识的完整性、科学性等；

最后由 AI 辅助优化表达方式与多媒体呈现，提升内容的吸引力与可读性。这种协作模式形成了有效的闭环验证，兼顾了效率与质量。与此同时，需针对不同受众群体（如青少年、专业人士）设计差异化的转译策略，例如对青少年强化类比与可视化，对专业人士则保留必要的技术细节与理论深度，实现知识的精准适配。

其二，强化生成内容的溯源能力与过程解释性。科学知识的可信赖性根植于其可验证性与透明度。因此，AI 系统在输出内容时，应清晰标注信息来源，并尽可能展示逻辑推理路径，便于用户追溯知识根源。进一步而言，运用可视化工具揭示 AI 的决策依据，有助于缓解“黑箱”疑虑<sup>[25]</sup>。例如，在处理存在科学争议的话题时，系统应明确标注不同学术观点的依据、证据强度及局限性，避免将单一视角呈现为绝对事实。这种做法不仅提升了知识呈现的多元性与辩证性，更有助于培育公众的批判性思维素养。

其三，实施分层级的知识表达与双向验证机制。科学知识本身存在层次化差异，从客观性的基本事实到探索性的前沿假说，AI 系统需具备识别并适配不同知识层级处理策略的能力，进而理清价值理念技术化的边界<sup>[26]</sup>。同时，引入“双向验证”流程很有必要，在将专业知识通俗化表达后，系统应能反向校验关键信息是否在简化过程中丢失或被曲解。这种双向校验机制是保障转译过程“保真”的关键，有效防止因过度简化而导致的认知偏差。

### 3.4 打开责任“黑箱”：锚定行动者角色与网络透明化

破解 AI 生成内容中的知识幻觉难题，关键在于厘清责任主体、提升技术透明度并构建协同治理框架。我们围绕责任界定、透明

度提升与协同共治 3 个维度展开，旨在解构责任“黑箱”，锚定各方角色，推动网络透明化，为 AI 知识传播的可靠性提供制度保障。

其一，构建权责明晰的责任分配框架与问责机制。AI 知识生产涉及多元主体共同参与，其中责任边界模糊造成治理效能弱化，甚至出现相互推诿、互相埋怨的局面。明晰责任边界是技术治理的基础性工作，一方面需要通过立法与行业规范共同完善，划定 AI 开发者、数据供应商、平台运营者及内容使用者在知识网络生态中的权责边界<sup>[27]</sup>。另一方面，AI 系统平台亟须建立内容标识制度，通过水印技术、元数据嵌入等技术手段增强内容溯源性，为责任认定提供实质依据。这种技术与制度的双重保障不仅有助于迅速锁定问题源头，更能在潜移默化中强化各主体的责任意识与自律行为。

其二，推进算法模型适度透明化发展。打开责任“黑箱”，意味着赋予算法模型新的“言说”能力，使其从“黑箱”转变为能够自我辩护的“透明行动者”。可解释性 AI ( Explainable Artificial Intelligence, XAI ) 技术将这种能力“铭刻”于模型之中。透明的算法行动者能够展示其决策路径，这迫使其他行动者（如开发者、监管者）无法再以技术复杂性为借口推卸责任，从而在网络中实现可见的问责机制。尽管商业机密与知识产权需得到保护，但这不应成为规避透明义务的理由。AI 技术服务商有责任向社会适度公开其模型的资料信息，如训练数据的情况、模型的能力边界等。特别是在科学传播等领域，应鼓励研发与应用可解释性 AI，向用户揭示决策逻辑，从而增强用户对 AI 生成知识的审辨能力与信任度。此举不仅有助于提升内容生成的稳定性，也为外部监管和公众监督提供了技术入口，有效遏制技术误用与信息操

纵的风险。

其三，构建多元主体协同参与的治理生态。鉴于 AI 知识幻觉问题的复杂性，单一主体规制方式力有不逮，亟须构建多元主体协同治理生态体系<sup>[28]</sup>。应建立由政府监管机构、技术企业、科研机构、科普机构、法律界与公众代表等共同参与的治理协商平台。具体而言，建立常态化的沟通与反馈机制，定期评估 AI 在科学传播领域的应用风险；联合制定行业伦理准则与实践指南；根据技术演进与社会反馈，对治理策略进行动态调整与完善。这种协同治理模式能够有效整合各方智慧与资源，提升决策的科学性，共同塑造更具韧性与可持续性的 AI 治理新生态<sup>[29]</sup>。

#### 4 结语

本研究基于行动者网络理论，系统揭示了 AI 知识幻觉的多维生成机制。研究表明，AI 知识幻觉并非单纯的技术缺陷，而是在复杂的行动者网络中，由于一系列关键“转译”环节的失败与背叛，导致网络整体目标偏离、信息传递失真的系统性后果。基于此，研究提出“四维协同”治理路径，即在源头网络构建多源数据验证与纯净性保障机制；在过程网络建立实时反馈与敏捷迭代系统；在转译网络开发语境感知型知识转化算法，强化人机协同的“保真”传递；在责任网络完善多元主体协同治理制度，明晰权责边界并推动透明化。本研究尝试探索了行动者网络理论从“对称性分析”到“非对称性治理”的应用转换路径，以回应复杂社会技术系统的现实需求。研究认为，构建可信赖的 AI 科普生态系统，本质上是技术持续优化与制度协同设计的动态演进过程，旨在促进异质行动者网络的良性互动，最终充分释放 AI 技术在科普中的应用潜能。

## 参考文献

- [1] 刘泽垣, 王鹏江, 等. 大语言模型的幻觉问题研究综述 [J]. 软件学报, 2025, 36(3): 1152-1185.
- [2] 坚持自立自强 突出应用导向 推动人工智能健康有序发展定 [N]. 光明日报, 2025-04-27(1).
- [3] Sun W, Shi Z, Gao S, et al. Contrastive Learning Reduces Hallucination in Conversations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13618-13626.
- [4] Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.
- [5] 何静, 沈阳, 谢润锋. 大语言模型幻觉现象的识别与优化 [J]. 计算机应用, 2025, 45(3): 709-714.
- [6] Ji Z, Lee N, Frieske R, et al. Survey of Hallucination in Natural Language Generation[J]. ACM Computing Surveys. 2023, 55(12): 1-38.
- [7] Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions[EB/OL]. [2025-06-01]. <https://arxiv.org/pdf/2311.05232.pdf>.
- [8] 何静, 沈阳, 谢润锋. 大语言模型幻觉现象的分类识别与优化研究 [J]. 计算机科学与探索, 2025, 19(5): 1295-1301.
- [9] Callon M. The Sociology of an Actor-Network: The Case of the Electric Vehicle[M]//Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology. London: Palgrave Macmillan, 1986: 19-34.
- [10] 吴莹, 卢雨霞, 陈家建, 等. 跟随行动者重组社会——读拉图尔的《重组社会: 行动者网络理论》[J]. 社会学研究, 2008(2): 218-234.
- [11] 全燕. 重组人机——行动者网络中的人机传播研究 [J]. 新闻与传播研究, 2023, 30(10): 39-51, 127.
- [12] 简圣宇. 生成式人工智能文艺创作的主体性问题 [J]. 上海师范大学学报 (哲学社会科学版), 2025, 54(1): 85-97.
- [13] 陈万球, 罗一人. 生成式人工智能的“知识幻觉”及其风险治理探论 [J]. 上海市社会主义学院学报, 2024(4): 38-51.
- [14] 陈昌凤. 人机何以共生: 传播的结构性变革与滞后的伦理观 [J]. 新闻与写作, 2022(10): 5-16.
- [15] 冯子轩. 生成式人工智能应用的伦理立场与治理之道: 以 ChatGPT 为例 [J]. 华东政法大学学报, 2024, 27(1): 61-71.
- [16] Callon M. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay[C]//Law J. Power, Action and Belief: A New Sociology of Knowledge? London: Routledge & Kegan Paul, 1986: 196-223.
- [17] 苏新宁, 吕先竟. 人工智能赋能人文社会科学研究方法变革 [J]. 西华大学学报 (哲学社会科学版), 2025, 44(1): 1-10, 121.
- [18] 袁曾. 生成式人工智能的责任能力研究 [J]. 东方法学, 2023(3): 18-33.
- [19] Mittelstadt B D, Allo P, Taddeo M, et al. The Ethics of Algorithms: Mapping the Debate[J]. Big Data & Society, 2016, 3(2): 1-21.
- [20] 陈忠. 形式主义的哲学实质、涂层物化与实践克服 [J]. 东岳论丛, 2025, 46(2): 51-57, 191.
- [21] Latour B. Reassembling the Social: An Introduction to Actor-Network-Theory[M]. New York: Oxford University Press, 2005, 39-40.
- [22] 王挺, 邵华胜, 王丽慧. 技术工具视角下的人工智能科普服务: 创新、风险与对策 [J]. 科普研究, 2024, 19(5): 5-13, 24.
- [23] 李猛. 深度合成技术的的社会安全风险: 样态表征、生成机理与敏捷治理 [J]. 中国科技论坛, 2024(5): 149-159.
- [24] 张燕翔, 朱育慧, 黄荣丽, 等. AIGC 语境下科普创作的科学性与叙事性协同优化策略研究 [J]. 科普研究, 2025, 20(2): 43-52, 80.
- [25] 胡泳. 人工智能驱动的虚假信息: 现在与未来 [J]. 南京社会科学, 2024(1): 96-109.
- [26] 闫宏秀, 李洋. 从价值对齐审视价值观技术化的有限性问题及其破解 [J]. 思想理论教育, 2025(5): 13-20.
- [27] 刁生富, 曹兰兰, 吴选红. 后真相时代深度伪造技术的信任问题与信任重构 [J]. 河南师范大学学报 (哲学社会科学版), 2023, 50(6): 80-85.
- [28] 王振波, 吴湘玲. 数字时代深度伪造技术研究——机理特征、功能异化及其优化理路 [J]. 北京航空航天大学学报 (社会科学版), 2025, 38(2): 47-55.
- [29] 冉连, 张薇. AIGC 中的深度伪造信息: 生成机理与治理策略——基于行动者网络理论的分析框架 [J]. 信息资源管理学报, 2025, 15(2): 137-150.

(编辑 颜 燕 和树美)