AI for Science中的机器幻觉：生成逻辑、伦理隐忧与治理模式

王 明[1]，张 文[2]
（1.东北财经大学 马克思主义学院，辽宁 大连 116025；
2.西安交通大学 马克思主义学院，陕西 西安 710049）

摘要：人工智能驱动的科学研究（AI for Science）正成为人工智能应用的重要领域之一，世界范围内与AI for Science相关的重大科学发现已证明其能够加速科学进程且具有强大发展潜力。然而，AI for Science中的机器幻觉正成为影响其强大潜力爆发与可持续发展的关键制约因素。AI for Science中的机器幻觉具有固有生成逻辑，包括数据偏差嵌入的认知幻觉、反馈机制固化的结构性谬误与推理均衡的脆弱性。同时，AI for Science中的机器幻觉也引发了多重伦理隐忧：科学研究结论的失真隐忧、科学研究严谨性的解构隐忧以及理论惰性的固化隐忧。为了推动AI for Science与人类价值观对齐，确保更好地服务于人类科学研究，现在就要探索AI for Science中机器幻觉的协同治理模式，构建逻辑一致性验证机制，推动技术维度的去幻觉化；重塑可验证AI科学体系，构建可信的制度维度；构建科研创新的伦理化机制，明确社会维度的责任导向；构建共治机制的制度，推动全球视野下的科学自治。

关键词：AI for Science；科学研究；机器幻觉；模型；风险
中图分类号：G301；TP18 **文献标识码**：A **DOI**：10.3969/j.issn.1003-8256.2026.04.004

中国科学院院士鄂维南表示：“人工智能驱动科研范式变革已成为科学界的共识”[1]。当下，以生成式人工智能为标志的新一轮人工智能技术不仅能够极大地扩展感知域与提升智力总量，而且能够与人类科学家进行深度协同，这些都能够加速科学研究发现进程，推动人类迈入科学发现第五阶段，即人工智能（AI）驱动的科学研究（AI for Science）范式。例如，2023年*Nature*发表题为Science discovery in the age of artificial intelligence的文章明确指出：“AI越来越多地帮助科学家加速研究，提出假设、设计实验、收集和解释大模型数据集，并获得了仅使用传统科学方法无法获得的见解”[2]。虽然AI在推动科学研究发现领域中具有巨大潜能，但即便是看似“聪明绝顶”的AI，也难逃“幻觉”的困扰[3]，而这些幻觉成为了限制AI for Science可持续发展的关键因素，导致AI for Science中的创造性幻觉特质与科学研究的严谨性之间存在着深刻的矛盾，这可能会引发科学研究结论的失真、科学研究严谨性的解构以及理论惰性的固化多重隐忧。

此外，由于AI for Science对于增强国家经济实力和国际竞争力具有重大意义[4]，这就意味着AI for Science影响力已经不再局限于科学研究单一领域，其影响力已经上升到国家战略层面。例如，2024年9月5日，美国能源部宣布投资6 800万美元资助11个项目，以促进AI在科学研究中的应用[5]。2025年8月26日，《国务院关于深入实施“人工智能+”行动的意见》提出要加快探索人工智能驱动的新型科研范式，加速“从0到1”重大科学发现进程[6]。《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出，以人工智能引领科研范式变革[7]。基于此，必须推动AI for Science可持续发展，使之更好地服务于科学研究，为人类创造更多福祉，现在就要从技术层面、社会层面、制度层面和全球层面着手探索协同治理模式。

1 AI for Science中的机器幻觉生成逻辑

AI for Science中的机器幻觉指模型在数据偏差嵌入、反馈机制固化与推理均衡脆弱性共同作用下，生成偏离规律、不可验证复现但表面自洽的伪科学性输出。AI for Science中的机器幻觉产生与模型训练过程密不可分[8]，其本质是模型预训练、训练与推理过程中多因素、多阶段共同作用的结果。数据采集、模型训练再到模型推理结果生成的贯通，使得AI for Science中的机器幻觉生成逻辑层层嵌套。首先，数据偏差嵌入的认知幻觉引发了模型训练的系统性隐患；其次，反馈机制固化的结构性谬误加剧了模型架构的内在缺陷；最后，推理均衡的脆弱性形成了解码策略的内在约束。

基金项目：国家社会科学基金项目（2020MYB038）；西安交通大学本科教学改革研究项目（2444Y）
作者简介：王明（1990—），女，甘肃天水人，讲师，硕士生导师，研究方向：人工智能安全与治理问题研究。E-mail：2186594619@qq.com

1.1　数据偏差嵌入的认知幻觉：模型训练的系统性隐患

“AI 的知识体系基本来源于训练时‘吞下’的数据源”[8]，这就使得作为驱动模型训练“燃料”的数据对于模型优劣具有重要影响，尤其是在数据采集环节出现的数据语料污染、采集领域覆盖面不足、采集过时数据等问题都可能使得模型由于数据性偏差而诱发事实性幻觉、忠实性幻觉以及价值对齐失败，这形成了 AI for Science 中模型训练的系统性隐患。

1.1.1　AI for Science 中数据采集的语料污染加剧科学研究中的认知幻觉

AI for Science 中数据采集中的语料污染风险指在模型预训练阶段采集到的用于科学研究的数据掺杂了低质量文本（携带误导性语义结构的数据、包含偏见内容甚至虚假信息等的数据），这些低质量文本可能在用于科学研究的模型正式训练过程中被其吸收学习，使得这些模型可能生成带有事实性幻觉的偏差性内容，甚至可能放大社会偏见与生成危害社会的言论。一方面，语料污染可能会破坏用于科学研究的模型对真实世界知识结构的学习能力。当前主流模型如 ChatGPT、Llama、Claude 等均依赖大规模地抓取互联网语料进行无监督式预训练，此种训练模式的突出特点是缺少事实核查环节，这可能会导致此类模型将未经验证的信息，如社交媒体谣言、虚构新闻等同于客观事实编码进科学研究的内容生成中，从而生成携带事实性幻觉的内容。另一方面，语料污染甚至还可能放大社会偏见与生成危害社会的言论，这会削弱用于科学研究模型的价值中立性，导致价值对齐失败。例如，一旦用于科学研究的训练语料中存在携带文化偏见、种族歧视、性别刻板印象与地域仇恨等内容时，AI for Science 的模型将以统计规律强化这些关联，从而生成放大偏见与生成危害社会的言论，从而加剧了模型训练的系统性隐患。例如，据美国《麻省理工科技评论》发表的一项国际研究指出：大语言模型（LLM）正悄无声息地传播全球各地的刻板印象，从性别歧视、文化偏见，到语言不平等，AI 正在把人类的“偏见行李”打包、升级，并以看似权威的方式输出到世界各地[9]。

1.1.2　AI for Science 中数据采集领域覆盖不足加剧科学研究中的认知幻觉

数据采集领域覆盖面不足也是诱发应用于科学研究的模型产生事实性幻觉的重要诱因，具体表现为当采集的数据在科学研究的专业领域覆盖面不足或代表性不强时，这些模型不能良好地形成对这些专业领域知识的结构化、稳定性与可靠性嵌入，这可能会导致该模型在专业问答任务中会优先通过语义填充或统计联想方式生成表面上看似合理实则缺乏客观依据的伪科学结论。特别是在医学等高精度与高风险的科学研究领域，语义表达要求专业性高，若当该领域采集到的训练数据未充分涵盖权威医学文献等时，用于驱动医学研究的模型便可能基于非专业语料进行类比性学习，这可能会导致借助该模型形成的科学研究结论发生错误归因或虚构药效等问题（如将保健品虚构为治疗性药物），从而加剧科学研究中的认知幻觉。更为严重的是，当采集到的用于科学研究的数据覆盖面不足时，还会破坏应用于科学研究中的模型可解释性与溯源能力。尤其是当模型用于执行开放式问答科学研究问题时，研究人员难以区分模型是否基于客观事实与真实数据做出推理，特别当模型生成内容表面看似严谨实则逻辑错误的内容时，易给科研人员形成“信任陷阱”，从而加剧科研人员的认知幻觉。例如，2023 年 2 月 3 日，荷兰阿姆斯特丹大学多位教授在 *Nature* 上发题为“ChatGPT: five priorities for research”的文章指出：“向 ChatGPT 提出需要深入理解文献的问题时，它经常给出错误和误导性的答案”[10]。

1.1.3　AI for Science 中数据采集的过时加剧科学研究中的认知幻觉

采集过时数据可能导致用于科学研究的模型训练时效滞后，这是此类模型产生事实性幻觉、价值对齐失败，增加模型训练的系统性隐患的又一重要原因，其加剧了用于科学研究模型的认知幻觉。以 ChatGPT-4 为例，其训练主要依赖静态快照数据集，如抓取 2023 年前互联网语料、旧版数据库等进行一次性预训练，这种训练模式决定了其知识更新的滞后性。当这样的模型用于科学研究时，其可能生成过时内容或已被证伪的内容，这会影响 AI for Science 的实用价值。例如，在医疗研究领域，要求模型采集最新的临床知识、新药上市信息，若模型仍基于旧版诊断标准或已被否定的药理机制进行推理，可能会造成误导性研究结论，这不但会影响医学科研人员声誉，更为严重的是威胁患者生命安全，增加了用于科学研究模型的不安全风险。

1.2　反馈机制固化的结构性局限：模型架构的内在缺陷

模型架构的内在缺陷是指源自模型设计本身的目标函数设计过于单一、过拟合与泛化失衡、人类反馈强化偏差等形成的模型结构性局限，模型即使在理想训练条件下也难以克服这些局限。

1.2.1　应用于科学研究的模型目标函数设计过于单一引发的模型架构内在缺陷

当下，用于驱动科学研究的绝大多数模型的目标函数设计采用优化目标的方式，如基于强化学习的用户奖励信号，这样的模型设计过分侧重提升人类评审者满意，从而导致忽视了对其他重要价值维度如可验证性和伦理约束等的提升，这可能会进一步强化反馈机制固化的结构性局限，进而增加诱发 AI for Science 中的忠实性幻觉概率。一方面，当应用于科学研究的模型优化目标仅停留于提升人类评审者满意度时，其更倾向于生成表面看似逻辑自洽但事实虚构的内容，以追求获取人类评审者较高评价分数，这也会增加诱发忠实性幻觉的概率。例如，2024 年 *Nature* 发表的名为“Larger and more instructable languagemodels become lessreliable”的文章

指出:“大参数模型不会承认它们的‘无知’,而更倾向于生成错误答案以赢得人类满意”[11]。另一方面,当应用于科学研究的模型目标函数缺乏多目标优化机制时也会削弱模型在知识边界意识和不确定性表达方面的能力。如果这样的模型不显示训练对认知盲区的识别能力,就难以学会在回答科学研究问题时引入诸如“尚无定论”“该问题仍存争议”等客观真实的表达,取而代之的是其可能会以貌似准确且确定的风格生成答案,这会抑制此类模型生成客观表达的机制,从而增加此类模型诱发忠实性幻觉的风险。

1.2.2 应用于科学研究的模型过拟合与泛化失衡引发的模型架构内在缺陷

模型过拟合与泛化失衡是模型在训练与推理阶段面临的核心挑战之一,尽管有些模型在大规模语料上表现出色,但当训练过程中出现过拟合特定模式、术语或语义结构时,其泛化能力会显著削弱,导致模型在处理新问题或复杂语境时生成语义扭曲、逻辑脱节甚至事实错误的内容。尤其是当模型推理能力越来越强时,其越可能将并无实际关系的数据潜在联系在一起,导致生成回答中专有概念“张冠李戴”,客观规律“拗曲作直”[12]。造成模型过拟合与泛化失衡并非简单的训练数据问题,而是模型结构性建模机制与目标函数设计的协同缺陷所致。过拟合导致模型在训练数据上学习过度,甚至捕捉到了训练语料中的偶然性、偏差性或其他噪声信息,导致应用于科学研究的模型丧失了对未知输入保持结构性理解及推理的能力,尤其是参数量超亿级的Transformer架构模型具有超强记忆能力,这使其易将训练中高频出现的关联模式固化为生成机制中的默认解释框架,从而形成在推理时以旧代新、以偏概全的认知错配,进而使得模型的泛化能力呈现出不足。例如,当这样的泛化能力不足的模型在处理量子计算对传统通信协议挑战任务时,其可能因泛化缺陷将量子计算错误地嵌套进牛顿力学中进行解释。而这样的解释内容表面看似语言流畅,实则存在跨学科的逻辑混乱错位,这可能会进一步加剧此类模型的忠实性幻觉。

1.2.3 人类反馈强化偏差引发的模型架构内在缺陷

人类反馈强化学习(RLHF)被用于提升模型对齐性能以改善用户体验。但是当前的RLHF机制存在固有缺陷,即更倾向于奖励那些语言表达流畅、貌似回答自信的模型生成,但与事实一致以及具备知识边界意识的生成内容通常被给予较低反馈评价分数。这种反馈机制可能会诱导用于驱动科学研究的模型形成修辞掩盖事实式的生成策略,从而形成算法固化的结构性谬误。一方面,人类评审者更倾向于对叙事流畅性偏好的模型生成给出更高评价分数,即那些表面看似语言组织良好、逻辑清晰的生成内容大概率会被人类评审者视为可信优质文本,这反过来也会强化这些被人类评审者标注高分的生成内容更具看似表达确定、自信甚至略显夸张的属性;另一方面,那些能够客观地表达不确定性、承认知识盲区的生成内容则被人类评审者给出较低评分。例如,Anthropic在研究经过RLHF训练的模型时发现:基于人类评审者偏好判断的训练方案可能会以不可取方式利用人类判断,如鼓励模型生成吸引人类评审者但实际上存在缺陷或错误的内容[13]。这种反馈机制可能会在无形中放大用于驱动科学研究的模型生成的表演成份,从而负面地影响模型策略梯度的更新方向,进而使得模型强化此种错误的生成策略,即与其承认不知道不如自洽地说错,最终加剧模型架构内在缺陷诱发忠实性幻觉。

1.3 推理均衡的脆弱性:解码策略的内在约束

模型推理解码策略局限指当前主流模型在生成推理过程(如数学推导、逻辑推断、科学解释等)时,受到其解码策略局限,导致生成内容在事实准确性、逻辑一致性或多步骤推理连贯性方面与事实忠实性之间难以维系平衡。近期,亚利桑那州立大学研究团队发布的一项题为《大语言模型的链式思维推理是海市蜃楼吗?数据分布视角的分析》的预印本研究指出,现有成果表明模型并非是“有原则的推理者”,而更像是“推理式文本的复杂模拟器”[14]。导致这一局限的关键原因在于模型训练过程中过分追求语言概率最优而非逻辑最优,意味着模型输出的“流畅性”并不等同于其内在逻辑的“健壮性”,这在AI for Science需要深度逻辑建模的研究领域形成了明显的瓶颈。

1.3.1 模型由概率驱动导致生成内容的随机性

模型主要学习的是语言分布,生成的内容是语言上常见的答案,而非逻辑上唯一正确答案。例如,GPT系列模型的核心原理在于通过深度学习海量文本数据,构建一个强大的概率模型[15],这就导致模型由概率驱动的生成内容中的“最可能”不等同于“最合理”。一方面,基于token概率采样的解码策略(如Beam Search)在自然语言生成任务中被广泛应用,然而在复杂知识表达与多步推理场景中,这种策略存在明显的局部最优偏置问题,即过度关注短期语言连贯性而忽视全局语义一致性与事实正确性。这一局限在生成如历史事件、科学推理或因果链条等要求高度结构一致性的科研文本时尤为突出。例如,当“嘉靖”“永乐”“乾隆”年号等多出现在类似上下文中,模型会基于语言频率和上下文相似度误判时间与人物的搭配,导致生成“乾隆年间,李世民巡视江南”等类似的逻辑上错误但语言上流畅的文本。另一方面,基于token概率采样的解码策略还存在路径偏倚问题,即倾向于生成过短或过长的句子,这在处理有固定结构和时间线的文本时可能会破坏原始事实的完整表达。

1.3.2 多模态对齐失败导致模型生成内容偏差甚至错误

在图文生成、匹配、问答等跨模态生成任务中,由于模型缺乏对真实世界实体、属性与关系正确的Grounding能力,从而导致多模态对齐失败,进而导致模

型生成时在常识、逻辑、规律等方面可能会出现偏差甚至错误。例如，在AI驱动的天文学研究领域，模型可能会将“太阳从西边升起”此种违背常识的结论与太阳升起的正常图像进行错误对应，导致语义与视觉脱节。一方面，这是由模态表征空间未对齐引起。当前主流模型采用对比学习或图文匹配损失函数在共享语义空间中对齐视觉和文本嵌入，但这种浅层嵌套对齐机制更关注全局语义相似性而非事实语义一致性，这就使得模型可能学到“太阳”“升起”“天空”等模糊关联特征，而未真正学会“东升西落”的物理规律，导致生成的内容虽表层图文相关但逻辑上错误配对。另一方面，现在的主流模型基本未显示整合世界知识库如ConceptNet等，这就会使得模型在生成时仅基于图像视觉token的相关性预测，导致模型在推理时忽视跨模态的因果性、不可逆性与物理规律约束。例如，虽然类似“太阳从西边升起”这样的表述在数据训练语料中极少出现，但是在缺乏物理规律约束的情况下模型依旧可能会生成，因为这样的表述在语言上句式上合理，且图像中若无强提示模型也无法纠正此错误。

1.3.3　长程依赖问题导致模型生成内容自相矛盾

“长程依赖问题指长文本中模型难以有效捕捉相距较远的词或句子之间的关联问题”[16]。在AI驱动科学研究实践中，尤其当模型处理科技报告、政策文件或长篇技术分析时，其若未能准确捕捉并维持跨句与跨段的逻辑一致性，极易导致生成内容出现前后语义矛盾、推理断裂或立场混乱等错误。例如，在同一份科技报告中，模型可能先论证某项技术具备可行性，随后又基于不同上下文逻辑否定这种技术的可行性，这是典型的长程依赖问题所致。一方面，这是由模型自注意力机制的理论局限所导致的。Transformer架构中的自注意力机制理论上应具备全局信息建模能力，但实际上由于其注意力稀疏化趋势使得随着模型所处理的文本长度增加，其更倾向于关注局部上下文而忽略远距离信息，从而使得自注意力机制在实际应用中全局信息建模能力不足。即使引入位置编码，模型也难以持续保持对跨段关系、话语立场、因果链条等长距离依赖的全局信息建模，从而导致生成内容在语义一致性与主题延续性上出现断裂。另一方面，这是由模型的生成策略局部最优偏置所导致。模型的多数解码依赖于局部token概率，但是由于其缺乏全局一致性约束机制，从而导致模型无法回溯先前生成内容的论证逻辑或立场，尤其在语言分布密集的任务处理中模型更容易被表层的语言流畅所诱导，从而忽略整体语义架构，进而在技术立场或预测结论层面诱发自相矛盾。

2　AI for Science 中的机器幻觉诱发的伦理隐忧

人工智能赋予科学研究新前景的同时，亦内隐多重潜在风险[17]，AI“幻觉”可能会引发严重后果[18]。本研究认为AI for Science中的机器幻觉诱发的伦理隐忧包括：科学研究结论的失真隐忧、科学研究严谨性的解构隐忧以及理论惰性的固化隐忧。

2.1　科学研究结论的失真隐忧：研究客观性的系统性侵蚀

OpenAI一篇名为“Why language models hallucinate”的论文得出结论：只要训练数据中不可避免地存在长尾、叙述和充满噪声的部分，模型在判断层面必然会失败。而模型在判断上犯的每一个错误都会被放大并传导至生成任务中，因此目前人工智能幻觉的局限性导致AI for Science中的机器幻觉不可避免，而其中对科学研究造成的最负面影响是科学结论的失真问题。

2.1.1　数据与事实不符导致科学结论失真

虽海量数据、前沿技术和强大计算能力貌似铜墙铁壁地保证着实证发现的信度，但却令人眼花缭乱的“数据技术密集型”实证研究得出的结论却常常显得“不靠谱”[19]。AI for Science驱动的科学研究中，数据与事实不符可能会导致科学研究结论失真，这是因为模型可能在执行科研任务时嵌入虚假实验数据与虚假参考文献，从而导致模型生成错误统计结果，而科研人员可能会在这些错误统计结果尚未验证的情况下将其当作正确统计结果并以此为研究基础推进实验，这可能会在很大程度上增加实验风险，尤其在药物研发等敏感科学研究领域更易引发科技伦理担忧，这可能会削弱公众对科学研究的信任。

2.1.2　推理与逻辑链断裂导致科学结论失真

在AI for Science驱动的科学研究中，推理与逻辑链条断裂可能会导致科学结论失真，这是由于模型在执行科学研究任务时，通常依赖概率驱动的语言模式匹配，从而易导致模型生成逻辑跳跃或因果关系误判等内容。例如，模型可能在仅发现变量间的统计相关性时便“武断”地将其认定为因果关系，这种以“相关性替代因果性”的模型推理错误，不但会破坏模型推理链条的逻辑完整性，而且还会使生成的科学研究结论在形式上看似自洽、符合学术表达规范，但实质上却建立在虚假或错误的推理路径之上，从而掩盖科学研究中潜在的混杂因素与复杂运行机制。

2.1.3　失真的科学结论可能会引发知识体系的系统性偏差

失真的科学结论一方面貌似合理自洽，另一方面具有隐蔽性，这就可能会导致科研人员在缺乏严格交叉验证的情况下易将这些失真的科学结论视为可信结论。长此以往，其还可能通过学术引用等途径逐渐扩散至更大范围科学共同体，导致研究客观性遭到系统性的侵蚀。因此，数据与事实不符、推理与逻辑链条断裂所诱发的科学结论失真风险，其不但挑战了科学研究中最为核心的因果推理标准，而且也动摇了科学研究结论的可

信性与可复现性。此问题的严峻性在医学等高风险领域尤为突出,因为失真的科学结论一经被采纳为临床依据,将直接转化为社会性风险。

2.2 科学研究严谨性的解构隐忧:可证伪性与复现性的缺失

“‘机器幻觉’正在成为制约AI科研可靠性的关键问题”[20],AI for Science的可靠性依赖于其可证伪性与复现性实现。一方面,科学研究的可证伪性意味其研究成果能够明确回溯至其数据来源、实验方法与逻辑推理路径;另一方面,科学研究的复现性意味着其研究结论可以在相同条件下被重复验证并得到相似结果。AI for Science中的机器幻觉则破坏了科学研究的这两大核心基石,从而可能诱发科学研究方法伪造、数据虚假等风险,进而加剧科学研究的学术严谨性解构隐忧。

2.2.1 AI for Science中的机器幻觉导致科学研究方法伪造

科学研究普遍强调研究方法的透明性、可验证性与可复现性。AI for Science中的机器幻觉可能导致模型生成的研究方法存在逻辑断裂等问题,而科研人员假若在未经验证情况下采用此种研究方法继续推进科学研究,可能会导致科学研究方法伪造的问题,从而增加了科学研究可证伪与复现的难度。例如,当AI在生成实验设计时虚构了某种不可行的实验条件或错误的技术参数,形成伪造的研究方法,而这种方法的偏差会使科研人员在后续进行复现实验时面临无法再现的困境。需要注意的是,这种研究方法的伪造并非传统学术意义上的不端主观造假,而是一种源于AI智能系统的非人为伪造,这就使得这样的方法伪造应用于科学研究时更具隐蔽性,导致科研人员更难以察觉,因此其会诱发更严重的学术严谨性解构风险。

2.2.2 AI for Science中的机器幻觉可能会导致虚构数据注入科学研究

科学研究的严谨性高度依赖于驱动研究数据的真实性与可重复性。由于AI没有真假概念,更没有输出内容与真实世界必须相符合的要求[21],这会导致其在生成数据时以统计拟合或语言生成的方式填补训练数据空白,从而导致模型生成看似合理却缺乏实际实验支撑的虚构数据。这类虚构数据往往具有较高的表观可信度,即它们在统计分布或形式上与真实实验数据高度相似,但其背后并无实验记录、测量过程或原始观测支撑,这会增加科学研究可证伪性与复现性的难度,从而加剧了科学研究严谨性的解构风险。

2.3 理论惰性的固化隐忧:创新性科学发现的抑制

常规意义上的机器幻觉生成呈随机性与无序性,而AI for Science中的机器幻觉以符合先验方式生成,即模型在生成知识或进行推理时倾向于迎合已有的假设框架、统计规律或学术共同体长期积累的主流范式。这种生成路径具有表面合理性且极具迷惑性,因为其能够使模型生成结果在逻辑或形式上与科研人员既有认知保持一致,这虽然更易获得科研人员信任,但是本质上这会增加理论惰性固化的隐忧,从而抑制创新性科学发现。

一方面,“使用AI进行科学创新和发现与人类努力的其他领域使用AI相比呈现出独特的挑战”[22]。用于驱动科学研究的模型先验迎合的生成方式会削弱科研人员怀疑、反驳与突破的科学精神,可能会使他们在面对模型生成时更倾向于接受与其既有假设一致的内容,这会导致科研人员忽略科学研究结论中的潜在异质内容。长此以往若这样的问题得不到有效解决,可能会进一步加剧科研人员的理论惰性,从而使科学研究陷入“回音室效应”困境,进而抑制创新性科学发现。

另一方面,AI for Science中的机器幻觉以符合先验的方式生成科学研究内容,这可能使科研人员将模型假设中的简化条件误当作科学真理,从而增加了理论惰性的固化风险。例如,在新药物发现或气候模拟等领域,模型通常依赖于高维复杂的现象降维表示和数学近似,而这些近似本身带有方法论上的限制与假设。当模型输出研究结论时,假若将这些简化条件以幻觉方式嵌入研究结果并以“科学新发现”方式面世可能会加剧理论惰性的固化风险。例如,*Nature*杂志发表的题为“Artificial intelligence and illusions of understanding in scientific research”的文章指出:“AI在科学研究中的广泛采用可能导致一种‘理解幻觉’的产生,即科学家可能误以为自己对研究对象有更深入的理解,而实际上这种理解可能是被AI工具简化或扭曲形成‘单一化知识生产’,这会使某些研究方法和问题的主导地位压制了其他创新方法的产生,从而降低科学的多样性和创造力[23]。”

3 AI for Science中的机器幻觉协同治理模式

AI for Science中的机器幻觉生成机制的复杂性与诱发的伦理风险的严峻性使得对其协同治理成为一项复杂性工程。因此,亟须人工智能企业、政府、社会和国际组织协同探索治理模式:构建逻辑一致性验证机制,推动技术维度的去幻觉化;重塑可验证AI科学体系,构建可信的制度维度;构建科研创新伦理化机制,明确社会维度的责任导向;构建共治机制制度,推动全球视野下的科学自治。

3.1 构建逻辑一致性验证机制,推动技术维度的去幻觉化

人类技术发展史反复验证了这样一条真理:技术应用引发的问题可以通过技术创新加以解决。AI for Science中的机器幻觉首先可以从技术层面通过开发去幻觉的逻辑验证工具加以攻克。逻辑验证工具是提升

模型可信度、防止科学结论失真、保障跨学科应用稳健性、支撑科研可复现的前提，此应成为协同治理路径中的“基础性”路径。

3.1.1　设计形式化验证方法

设计形式化验证方法不是“黑盒测试”的增强版，而是把科学问题的不变量与推理规则以数学逻辑形式表现出来的方法，旨在提供模型可检查的正确性证明[24]。设计形式化验证方法需做好以下三方面工作。一是模型的规则提炼与形式化。这要求负责开发模型的技术人员不仅要与 AI for Science 领域专家确认模型须遵守的规则，如变量单位一致、因果关系不倒置及能量守恒等，而且还需将这些规则转化为模型可读的数学表达式或逻辑公式，如不等式与时序逻辑等以确保逻辑验证工具可直接理解与执行。二是模型的系统抽象与验证实现。这需要技术人员不仅要将模型拆分成可分析的模块，并为每个模块建立可处理的逻辑或数学模型，而且还需要技术人员选择合适验证工具运行验证，从而判断模型是否存在违反规则的输入或逻辑链条。三是反例处理与闭环改进。若技术人员在验证中发现模型输出异常，需分析导致这些异常产生的原因到底是规则设定问题还是模型训练缺陷问题或是抽象过粗问题，且需要针对存在的这些异常问题再进行模型修复并再次验证，从而形成自动化闭环流程，进而使模型每次更新都能经过验证，最终确保模型输出始终符合 AI for Science 的领域专家要求。

3.1.2　构建动态监测框架

构建动态监测框架需要开发实时监测工具以跟踪模型在运行中的逻辑一致性、识别并拦截异常推理路径，需做好以下两方面工作。一方面，检测实时逻辑一致性。(1)确定要检测什么(What)，这包括对数学不变量(能量守恒、质量守恒)、单位维度一致性、变量边界与单调性、因果顺序、数值稳定性等的检测；(2)确定如何检测(How)，这包括在推理管线拦截中间态与最终输出使用轻量规则引擎+数值残差计算。值得注意的是，对高风险任务采用阻断模式，一般任务用非阻断打标，避免增加显著延迟。另一方面，进行轻量日志分析与异常检测。(1)要记录模型结构化日志(What)，包括应用的模型版本、输入摘要、置信分布、关键中间量、时间戳与环境信息；(2)要建立流式管道(How)，包括数据采集、特征提取、打分、告警等环节；(3)处理好工程细节，例如保持日志样本(审计)并周期性回放用于模型回溯。

3.1.3　构建多模态数据源交叉验证机制

多模态数据源交叉验证机制通过引入图像、实验数据、传感器信号、数值仿真结果等异源信息，为 AI for Science 中机器幻觉诱发的伦理风险治理提供可溯源证据链，使得每个模型生成结论都能被不同模态的独立证据支撑或驳斥，从而助力提高模型生成结论可靠性。例如，面对 Deepfake(深度伪造)技术已能以极低成本生成高度逼真的图像、音频和视频等，传统的单一模态检测手段已难以应对，因此需构建多模态数据源交叉验证机制以提升内容真实性审核能力。构建多模态数据源交叉验证机制需做好以下三方面工作。(1)构建可信度高的异源模态体系与语义对齐框架。这不仅需要针对科学研究的目标问题系统性编制模态目录，而且还需要明确各模态在证据链中的量化特征、误差分布、先验可信度与逻辑地位。(2)设计冲突信号检测机制。当模型生成内容的冲突信号超过阈值时，这样的检测机制应能够自动触发人工复核程序，并能够将该生成内容标记为幻觉产物。(3)构建人工介入、纠错迭代与审计可追溯的治理闭环机制。这要求建立对 AI for Science 中的机器幻觉的分级处置机制，此机制应能够根据一致性检验结果将模型生成内容划分为自动接受、自动拒绝及人工复核三种类别。其中对于自动拒绝和经人工复核生成的幻觉内容应建立错误样本数据库，并按如数据噪声、模态缺失、模型推理偏差等幻觉诱因将错误样本进行分类，以便驱动模型提质再训练。

3.2　重塑可验证 AI 科学体系，构建可信的制度维度

正如清华大学互联网治理研究中心学术委员会主席薛澜教授在 2025 中国数字经济发展和治理学术年会发表的题为《全球视野下的人工智能治理——挑战、机制与未来路径》的主旨演讲所认为的那样：“人工智能治理的目的，不仅是‘应对风险’，更重要的是塑造人工智能的社会-技术系统(socio-technical systems)、支撑相应的制度架构的协同演化，并最终形成一个兼具创新性、安全性与公平性的社会应用生态”[25]。当下各国政府对人工智能应用风险治理多采用事后追责的治理方式，而随着人工智能普及性应用其诱发的应用风险更加广泛多样，特别是在 AI for Science 中，事后追责已无法满足这种技术的应用需求。因此，在 AI for Science 实践中，应做好风险的前置性预防工作，这就需要在制度维度设计层面重塑可验证 AI 科学体系。

3.2.1　对模型合规性进行认证

模型合规性的认证是确保 AI 驱动科学研究可验证的基础。做好此项工作需从以下三方面着手。一方面，对模型进行严格的算法透明度测试。此类测试要求模型在结构化验证框架下能够系统解释其内部推理与决策的核心环节，并且这样的解释性应符合 L1 级别的标准①。这种较为严格的测试要求意味着模型须具备高精度因果链解析能力、不确定性标注能力以及可重构推理轨迹存储能力。另一方面，向独立第三方提交涵盖模型训练所使用的全部数据来源完整清单，此清单须包括对全部数据的出处、标签、获取授权情况、数据质量评估结果以及数据偏差风险点进行明确的标注。

① L1 级别标准是指模型可解释性不仅限于表层的输入-输出关联描述，而应在特征贡献度、参数交互效应与逻辑链条等多维度实现可追溯，模型至少能够还原并解释 80% 以上的关键决策路径。

3.2.2 设计风险防控预案

对于AI驱动的高敏感高风险领域的科学研究而言尤其需要设计风险防控预案，这有助于减少不可逆的重大科学误导或重大社会风险发生。一方面，诸如AI驱动的医学研究等高敏感高风险研究领域，此类科学研究中的模型除需满足可解释、透明等基本要求外，还需额外基于多维度指标构建动态风险评估测试机制。这样的测试要求模型在仿真与真实应用场景中能够面对多类型对抗样本攻击（如梯度导向扰动、分布漂移注入、数据投毒及逻辑推断攻击等）时保持较好的鲁棒性，并且这样的模型还能够通过跨时间窗口的压力测试与输入扰动敏感性分析。而经过风险测试的合规性模型评估结果应在上述攻击条件下稳定性误差率控制在0.5%以下。另一方面，用于驱动科学研究的模型还需内置实时异常中断机制。此类中断机制旨在确保当检测到模型在推理过程中发生异常情况时能够立即中止当前计算进程并触发多层级响应程序，包括本地隔离、审计日志生成与人工复核介入。

3.2.3 对驱动科学研究的模型形成评估报告

做好此项工作需从以下三方面着手。一方面，模型开发者要形成对此类模型的评估报告。中国科学院院士鄂维南认为："AI for Science涉及的应用场景，包括工业制造、材料、药物等都是实体经济不可或缺的"[26]。党的二十届四中全会提出"全面实施'人工智能+'行动，以人工智能引领科研范式变革，加强人工智能同产业发展、文化建设、民生保障、社会治理相结合"。由此可见，人工智能引领的科研范式变革与人类生产生活日益密切且影响愈加深远广泛。因此，对用于驱动科学研究的模型对社会可能造成的负面影响需要前置性评估，这就要求由模型开发者基于跨学科方法，系统性地分析此类模型在预期与非预期应用情境下可能对社会产生的负面潜在影响。这些负面潜在影响应包括可能对公众认知、就业结构调整、资源配置以及弱势群体造成的负面冲击等。另一方面，对驱动科学研究的模型形成评估报告须跨学科专家小组审核通过。"当前开放科学运动蓬勃发展，全球科研模式正在发生深刻变革"[27]，跨学科研究已成为当前科学研究的重要范式。因此，这样的评估报告应邀请来自人工智能、社会学、法学、伦理学等不同领域的专家学者对评估报告的完整性与科学性进行逐条审查，尤其需要重点评估模型的风险识别能力、缓释措施的可行性等。

3.3 构建科研创新伦理化机制，明确社会维度的责任导向

人工智能的研发过程就是一部不断巩固、放大和提升人类主体性的历史，其在未来的每一点进步，都是对人本质力量的一次确认[28]。"从长期来看，人工智能的发展需要充分考虑可持续发展目标、推动创新创造，人工智能的开发使用者应当主动承担社会责任，让技术应用服务更为先进的生产力"[29]。因此，人与AI协同合作开展科学研究是当下及未来AI for Science发展的必由之路[30]。基于此，在社会维度的责任导向层面需构建科研创新的伦理化机制。

3.3.1 构建将"以人为本"伦理要求嵌入模型开发全生命周期机制

华南师范大学闫坤如教授[31]认为将"伦理先行"嵌入技术全生命周期，有助于确保人工智能健康发展。在AI驱动科学研究的实践中，在模型开发全生命周期中嵌入"以人为本"伦理要义，对于增加模型价值对齐与健康发展、推动AI for Science发展与人类福祉深度融合具有重要意义。构建这样的机制应覆盖模型从前期需求分析到后期运行维护的各环节，具体需要做好以下三方面工作。一方面，模型需求分析阶段应引入价值对齐测试。"价值对齐旨在让大模型的能力、行为与人类的真实意图、价值观以及社会道德准则相一致"[32]。模型需求分析阶段引入价值对齐测试需汇聚多学科专家共识建模，并对形成的模型进行多元文化视角评估，旨在对模型用于驱动科学研究时可能诱发的伦理冲突进行前瞻性情景模拟，以提前制定风险缓释方案，增强AI for Science目标与法律规范、社会伦理及公共利益的一致性。另一方面，模型训练阶段需建立高精度训练数据集偏见检测与修正机制。这需相关技术人员识别训练数据中存在的潜在系统性偏见与隐性歧视语料，并通过数据清洗、加权重采样等方式降低数据偏差对模型的影响。此外，模型输出生成阶段应配置多层级真实性验证模块。这需采用多模态交叉验证、事实核查引擎与实时语义一致性检测等方法，旨在减少模型生成断章取义、误导甚至虚假内容。需要说明的是，此机制建成后也要保持持续迭代与更新能力，应在模型部署后的运行周期内构建在线监测与用户反馈闭环机制，以实时跟踪模型在不同情境下的表现，方便相关技术人员依据模型诱发的新问题及时调整算法与审查标准。

3.3.2 制定AI for Science的行业自律标准

AI for Science的行业自律标准是构建负责任创新科研机制的内生约束机制，应在遵循"多方参与-证据-支撑-风险分级"原则指导下进行，确保制定的行业自律标准既能保障科学研究的严谨性又能确保其顺利落地执行。通常而言，制定AI for Science的行业自律标准需做好以下四方面工作。一是成立由科研机构、学术期刊、资助机构、产业界和公众代表组成的跨学科、跨领域AI for Science行业自律标准制定联盟。该联盟可以通过研讨会等形式达成自律标准共识，具体需明确行业自律标准的核心概念、适用范围以及最低伦理底线。二是需为AI for Science全流程建立清晰的规范。首先，在数据采集环节要遵循"可查找、可获取、可互操作、可重用"原则；其次，在模型训练环节要对模型能否提供解释等进行说明；最后，在科学研究成果的核验环节要对科学

研究成果进行事实核查并对其可能产生的社会影响进行评估。三是对 AI for Science 诱发的伦理风险进行分级管理。应按 AI for Science 用途与潜在负面社会影响将 AI for Science 项目划分为不同风险等级管理，尤其对被划分为高风险研究领域如医学项目需进行预注册、伦理前置审查与独立第三方机构复核。四是制定的 AI for Science 行业自律标准应及时更新并与国际标准如 ISO、OECD 等对接，以确保这些自律标准能持续适应 AI for Science 可持续发展的要求。

3.4　构建共治机制和制度，推动全球视野下的科学自治

“人工智能驱动的科学研究关涉人类命运共同体，需要全球通力合作，构建全球共识的道德标准，推动人工智能向善发展”[33]，而构建科学共同体自治体系是构建全球共识的道德标准的前提与基础，其能够把 AI for Science 的道德期待转化为工程化与制度化的可执行承诺，从而平衡科学前沿探索力与实现可持续发展间的关系。值得注意的是，这里的科学共同体自治体系中的“自治”不是指与政府监管对立，而是强调由跨学科、跨国界的科学共同体主导构建的 AI for Science 自治规则，其能够与技术层面、制度层面和社会层面的协同治理路径形成互补的“共治体”。

3.4.1　构建明确的科研规则和“科研通行证”制度

此类规则和制度旨在减少 AI for Science 诱发“黑箱化”风险，旨在推动 AI for Science 成果可验证、能复现与能信任。构建此类规则与制度应从以下三方面着手。一是制定统一且方便机器读取的标准化元数据模板。这需要记录模型训练所使用的全部数据来源，包含数据采集方式、数据质量评估与数据合法性证明等，此外还应包括模型架构、算力消耗以及版本控制与再现性验证信息等，增加 AI for Science 过程可追溯、研究结果可复现与跨领域可验证的概率。二是针对高风险领域的科学研究活动，如基因编辑、病原体合成等需建立分级管理体系，将相关数据和研究成果按风险类别划分为完全公开、延迟披露或受控访问等，并配套用途审查、访问许可与持续监测机制，确保 AI 在驱动科学研究的同时又能够有效防范技术滥用与风险外溢。三是需组建由跨学科领域同行专家与独立的第三方机构共同构成的常设审核团队，这不仅要依据预先设定的规则与安全评估标准对科学研究项目实施随机化抽查与系统性审计，而且还需要根据审查结果动态颁发、续期或撤销“科研通行证”，帮助科学研究活动向更加符合学术规范、伦理要求与风险控制标准方向靠拢，从而实现 AI for Science 的过程性监管与即时纠偏。

3.4.2　打造 AI“联合科学家”

打造 AI“联合科学家”的目标在于使全球科研人员能够在同一个安全、可复现的在线实验平台上开展科学研究工作，其突出优势是减少科学研究中的重复性工作，以降低研究成本，提高研究效率，增强正确科学研究成果的全球共享性。例如，2024 年美国斯坦福大学病理学家托马斯·蒙廷开启了一场前所未有的“实验室会议”，即与 6 位由人工智能（AI）驱动的“虚拟科学家”共商阿尔茨海默病的治疗策略，这些 AI 被赋予不同的专业角色，从神经科学家到药物化学家，在几分钟内展开多轮讨论，最终生成了一份长达一万多字的会议纪要。2025 年 2 月，谷歌旗下“深度思维”公司也推出了一款名为“AI 联合科学家”的软件，该软件由 6 个 AI 代理组成，分别负责想法生成、反思或批评、概念演进、去重、排序和总结审稿，均由谷歌的 Gemini 2.0 模型驱动。这套系统是谷歌生物医学 AI 研究工作的延伸。在一项早期测试中，该系统在两天内就解决了困扰科学家十多年的科学谜题。

3.4.3　构建奖励良好科研行为的激励机制

短期来看，推动科研数据共享、遵循科研伦理规范、复现科学研究成果等科研行为对科研人员的回报性价比不高，但是构建奖励良好科研行为的激励机制能够帮助科研人员让这些科研行为成为他们在科研实践中的加分项，从而进一步推动科研人员主动开展更透明、更负责任的科学研究，进而营造推动整体科研质量提升的学术氛围。构建奖励良好科研行为的激励机制的举措如下。一方面，需构建面向科研人员的量化信誉积分机制。这需将科研人员共享高质量数据与代码资源，以及在潜在伦理风险出现前主动提交预警报告等科研行为纳入激励机制的加分项，促进学术生态的长期健康发展；另一方面，鼓励科研人员公开发表负面科学研究结果与无效实验清单。这些负面科学研究结果与无效实验清单可以被纳入正规学术交流与评价体系中，且还可以通过设立专门期刊栏目对其进行学术探讨，以减少科学研究活动中只报告研究的成功案例而导致的知识偏倚，从而增强科学研究知识库的真实性与完整性。

参考文献：

[1] 新华网. 聚焦“十五五”规划建议丨以人工智能引领科研范式变革[EB/OL].(2025-11-19)[2025-11-23].https://www.news.cn/20251119/210ca51e9b6d42febf9889aea4d2e36e/c.html.

[2] WANG H C, FU T F, DU Y Q, et al.Scientific discovery in the age of artificial intelligence[J].Nature, 2023, 620(7972): 47-60.

[3] 刘霞. 生成式 AI“幻觉”困境如何破解[N]. 科技日报, 2025-01-28(4).

[4] 李建会, 杨宁. AI for Science: 科学研究范式的新革命[J]. 广东社会科学, 2023, 40(6): 81-92.

[5] 美能源部投资 6800 万美元促进 AI 在科学研究中的应用[EB/OL].(2024-10-10)[2025-08-01].https://ecas.cas.cn/xxkw/kbcd/201115_146304/ml/xxhcxyyyal/202410/t20241010_5035564.html.

[6] 中华人民共和国中央人民政府. 国务院关于深入实施“人工智能+”行动的意见[EB/OL].(2025-08-26)[2025-08-28].https://www.gov.cn/zhengce/content/202508/content_7037861.htm.

[7] 中华人民共和国商务部. 中共中央关于制定国民经济和社会发展第十五个五年规划的建议[EB/OL].(2025-10-28)[2025-11-23]. https://www. mofcom. gov. cn / xwfb / rcxwfb / art / 2025 / art_411d31b47a014eaaaee3d3f847b19b83.html.

[8] 徐剑.人工智能为何会产生幻觉(唠"科")[N].人民日报,2025-06-21(6).
[9] 张佳欣.AI输出"偏见",人类能否信任它的"三观"?[N].科技日报,2025-07-17(4).
[10] VAN DIS E A M, BOLLEN J, ZUIDEMA W, et al. ChatGPT: five priorities for research[J].Nature,2023,614(7947):224-226.
[11] ZHOU L X, SCHELLAERT W, MARTÍNEZ-PLUMED F, et al. Larger and more instructable language models become less reliable [J].Nature,2024,634(3):61-68.
[12] 陈紫璇,成祖明.大语言模型对学术研究的机遇与挑战[EB/OL].(2025-03-27)[2025-09-06].https://www.cssn.cn/skgz/bwyc/202503/t20250327_5864557.shtml.
[13] SHARMA M, TONG M, KORBAK T. Towards understanding sycophancy in language models [PP/OL]. ArXIV (2023-10-20) [2025-11-30].https://arxiv.org/abs/2310.13548.
[14] AI大脑"推理能力"被质疑:亚利桑那州立大学揭示链式思维的真面目[EB/OL].(2025-08-11)[2025-09-06].https://news.qq.com/rain/a/20250811A065C400.
[15] 大模型原理:GPT如何通过概率预测文本[EB/OL].(2025-08-09)[2025-09-06].https://blog.csdn.net/hzcbd/article/details/150109217.
[16] 语言模型中的长程依赖问题[EB/OL].(2024-08-10)[2025-09-06].bbs.huaweicloud.com/blogs/432631.
[17] 徐东波.人工智能驱动科学研究的逻辑、风险及其治理[J].中国科技论坛,2024,40(5):120-129.
[18] 叶菲楠,李侠.预测加工视角下的AI幻觉:本质、机制与治理[J/OL].科学与管理(2026-04-16)[2026-05-08].https://link.cnki.net/urlid/37.1020.G3.20260415.1059.006.
[19] 庞珣.避免"下不该下的结论":社会科学研究中的识别与信度[J].中国社会科学评价,2021,7(3):62-71.
[20] 邱慧丽.人工智能赋能科学研究要防范"机器幻觉"[N].学习时报,2025-09-03(6).
[21] 刘永谋.警惕AI"幻觉"带来的安全风险[N].中国科学报,2025-02-24(3).
[22] 杜严勇.人工智能驱动科学研究的机遇、挑战与对策[J].科学与管理,2025,45(2):1-7,103.
[23] MESSERI L, CROCKETT M J. Artificial intelligence and illusions of understanding in scientific research[J].Nature,2024,627(8002):49-58.
[24] 卜磊,陈立前,董云卫,等.人工智能系统的形式化验证技术研究进展与趋势[EB/OL].(2020-11-09)[2025-08-15].https://faculty.sist.shanghaitech.edu.cn/faculty/songfu/publications/CCF-report-2019-2020.pdf.
[25] 薛澜.全球视野下的人工智能治理:挑战、机制与未来路径[EB/OL].(2025-07-19)[2025-11-23].https://news.qq.com/rain/a/20250719A06U8L00.
[26] AI for Science:科学研究新范式[N].环球杂志,2023-05-10(9).
[27] 陆成宽.推进开放科学运动 构建高端学术交流平台[N].科技日报,2021-11-22(1).
[28] 张劲松.人是机器的尺度:论人工智能与人类主体性[J].自然辩证法研究,2017,33(1):49-54.
[29] 傅宏宇.构建"以人为本"的人工智能治理体系[N].学习时报,2024-03-22(A3).
[30] 于金龙.人工智能驱动科研的哲学审视[J].北京航空航天大学学报(社会科学版),2024,37(5):69-75.
[31] 闫坤如.人工智能伦理风险及其治理[N].学习时报,2025-11-19(6).
[32] 李思雯.人工智能价值对齐的路径探析[J].伦理学研究,2024,23(5):99-108.
[33] 李伦,刘梦迪.人工智能驱动的科学研究范式革命:态势与未来[J].探索与争鸣,2024,40(10):143-151,180.

Machine Hallucinations in AI for Science: Generative Logic, Ethical Concerns and Governance Models

WANG Ming[1], ZHANG Wen[2]
(1.School of Marxism, Dongbei University of Financeand Economics, Dalian 116025, China;
2.School of Marxism, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: The domain of artificial intelligence-driven scientific research, otherwise termed AI for Science, is emerging as a significant application domain for artificial intelligence. A plethora of significant scientific discoveries pertaining to AI for Science have been made on a global scale. These discoveries have demonstrated the ability of AI to accelerate scientific progress and its immense developmental potential. However, machine hallucinations within AI for Science are emerging as a critical constraint affecting the realization of its immense potential and sustainable development. These hallucinations possess inherent generative logic, including cognitive illusions embedded with data biases, structural fallacies solidified by feedback mechanisms, and vulnerabilities in reasoning equilibrium. Concurrently, machine hallucinations in AI for Science give rise to numerous ethical concerns, including the distortion of scientific research conclusions, the deconstruction of scientific rigour, and the entrenchment of theoretical inertia. In order to align AI for Science with human values and ensure it better serves scientific research, the exploration of collaborative governance models for machine illusions in AI for Science is now required. The following three-pronged approach is proposed, The establishment of logical consistency verification mechanisms to advance technical de-illusioning; The reshaping of verifiable AI scientific systems to build trustworthy institutional frameworks; The development of ethical mechanisms for scientific innovation to clarify societal responsibility; The construction of co-governance institutions to promote scientific autonomy within a global perspective.

Keywords: AI for Science; scientific research; machine hallucinations; models; risks