■ 教育数智化

DOI:10.15998/j.cnki.issn2097-6763.2026.03.006

第五科研范式下 AI 幻觉的生成逻辑、进阶样态与治理路径

胡金艳[1],董　澳[1],苏林猛[2]

(1.河南师范大学　教育学部,新乡　453007;2.河南师范大学　政治与公共管理学院,新乡　453007)

摘　要:人工智能驱动的科学研究(AI4S)作为第五科研范式正深刻地影响着科研工作。AI 生成看似合理但实际偏离客观事实的幻觉现象严重威胁科研可靠性,成为第五科研范式下的核心治理难题,但同时 AI 幻觉也蕴含着激发突破性创新的潜能。现有研究多聚焦技术纠错与算法优化,缺乏对 AI 幻觉生成逻辑、风险样态与治理路径的探讨。从技术哲学、协同学及计算认知神经科学等多维视角,揭示 AI 幻觉并非技术缺陷,而是人机共生系统中不可消除的结构性副产品。AI 幻觉呈现从显性"失序"、半隐性"有序的失序",到深度隐性"伪秩序"的渐进式三重演化样态,这一演化过程伴随着主体间性的动态涌现风险与认知层级的渗透性侵蚀风险。为应对此挑战,构建了以审辨式共生为核心理念的"三轮协同"治理模型,该模型涵盖动态权责系统、人机交互协议及协同治理基座,旨在确立研究者在人机共生中的主导地位,为转化 AI 幻觉风险、驾驭其创新潜能提供理论框架与实践路径。

关键词:人工智能;第五科研范式;AI 幻觉;审辨式共生;创新潜能

[**中图分类号**]G640;G301;TP18　[**文献标志码**]A　[**文章编号**]20976763(2026)03005212

修回日期:20260107

基金项目:国家社会科学基金一般项目"人智共生视角下人工智能'幻觉'的风险识别及韧性治理研究"(25BKX020)

作者简介:胡金艳,女,河南漯河人,河南师范大学教育学部副教授,博士生导师,教育学博士,主要从事人工智能+教育和学习科学与技术研究;
董澳,男,河南南阳人,河南师范大学教育学部硕士生,主要从事知识建构研究;
苏林猛,男,河南许昌人,河南师范大学政治与公共管理学院博士后,主要从事教育数字化、课堂教学行为分析和技术哲学研究。

引用格式:胡金艳,董澳,苏林猛. 第五科研范式下 AI 幻觉的生成逻辑、进阶样态与治理路径[J]. 重庆高教研究,2026,14(3):52-63.

Citation format: Hu Jinyan, Dong Ao, Su Linmeng. The generation logic, advanced patterns, and governance pathways of AI hallucination under the fifth scientific research paradigm[J]. Chongqing Higher Education Research,2026,14(3):52-63.

一、问题提出

2007 年,图灵奖得主吉姆·格雷(Jim Gray)将科学研究的范式分为 4 种:基于实验描述经验事实的经验范式、利用数学模型进行理论推演的理论范式、依托计算机仿真模拟复杂过程的计算范式,以及大数据时代基于海量数据挖掘发现规律的数据密集型范式。然而,随着 AI 技术的爆发式增长,人类社会正加速迈向人机协同的智能新时代,数据密集型范式越来越难以应对当下日益凸显的数据不确定性、复杂性、维数爆炸和尺度边界等问题,导致科学研究范式的深刻变革[1]。因此,有学者将人工智能驱动的科学研究(AI for Science,AI4S)视为继经验范式、理论范式、计算范式和数据密集型范式之后的第五科研范式[2-3]。与前 4 种范式相比,AI4S 不仅代表科研工具效率的升级,更标志着科学发现底层逻辑的重构,具体表现为:(1)AI 从辅助工具转变为深度参与科研核心环节的搭档[4-5],促使人机关系发生根本性变革,甚至有研究者将人机视为共生合一的新型主体[6-7];(2)打破传统线性研究路径,构建“假设生成—审辨识别—验证迭代”的智能闭环[8]。AI4S 的革新性潜力已在生命、材料、数学等多个学科领域得到证实[9],如学者通过机器学习探索大空间的单层过渡金属卤化物,实现半监督学习的机器学习形式,加速新材料的发现及其磁性的预测[10]。正因如此,近年来多国或国际组织将 AI4S 视为新科研形态,欧盟于 2023 年强调 AI 融入科学有助于在全球科学领域获得竞争优势[11]。中国科技部同年启动“人工智能驱动的科学研究”专项,布局 AI4S 前沿科技研发体系[12]。2025 年英国托尼·布莱尔全球变革研究所(Tony Blair Institute for Global Change)提出将 AI 深度融入科研体系[13]。可见,AI4S 作为第五科研范式正深刻地影响着人类的科研工作。

然而,随着 AI 技术深度介入高校科研工作,AI 生成错误或虚构信息的幻觉现象(简称“AI 幻觉”)成为科研探索中的核心治理难题。AI 幻觉并非第五科研范式的主动产物,而是技术深度嵌入智能闭环步骤引发的衍生风险。如研究者在使用 AI 搜索文献综述时,大模型为了迎合其意图而生成作者、年份、刊名俱全但实则不存在的参考文献;AI 辅助分析实验数据时,大模型生成真假参半的代码等[14]。这不仅对科研严谨性和可信度带来威胁,还可能引发科研人员认知外包和认知堕化等深层风险[15],冲击学术诚信体系。有研究者从 AI 幻觉的技术机制层面进行了探讨,并尝试语义校验等缓解策略[16],为后续研究奠定基础。但已有研究多将 AI 幻觉视为静态、单点的模型偏差,忽略了其在人机动态交互中的多维属性[17],且关于 AI 幻觉现象对科研诚信、知识创新的深层影响探讨不足[18]。AI 幻觉不仅是孤立的技术偏差,更涉及人机持续互动过程中技术、人类认知与知识生产关系的深层问题。因此,本研究拟突破静态、技术的分析框架[19],从技术哲学、协同学、计算认知神经科学的视角,探讨 AI4S 范式下 AI 幻觉现象的生成逻辑、风险样态及治理路径。本研究聚焦 3 个问题:一是探究 AI4S 范式下 AI 幻觉的生成逻辑;二是识别和分析 AI 幻觉对科研过程和成果可靠性构成的风险样态;三是构建科研人员应对 AI 幻觉挑战的有效治理路径。

二、多维理论视角下 AI 幻觉的生成逻辑

AI 幻觉并非智能技术演进中的偶发性缺陷,而是技术运行与社会应用过程中的副产品,其本质应被理解为一个多维度、跨层次的动态过程[20]。借鉴罗伊·巴斯卡(Roy Bhaskar)关于现实分层的洞见:要理解一个复杂现象,必须穿透其表层“经验域”“实际域”,潜入深层的“真实域”[21],本研究首先从宏观技术哲学视角回答其“为何必然存在”,揭示其作为技术药理的内在属性;其次,从中观协同学视角阐释其“如何动态发生”,呈现人机耦合系统的涌现规律;最后,从微观计算认知神经科学视角解析其“怎样实践显现”,捕捉算法层面的仿生偏差。三者分别对应 AI 幻觉的存在逻辑、发生逻辑与实践逻辑,此 3 个维度的逻辑框架如图 1 所示。

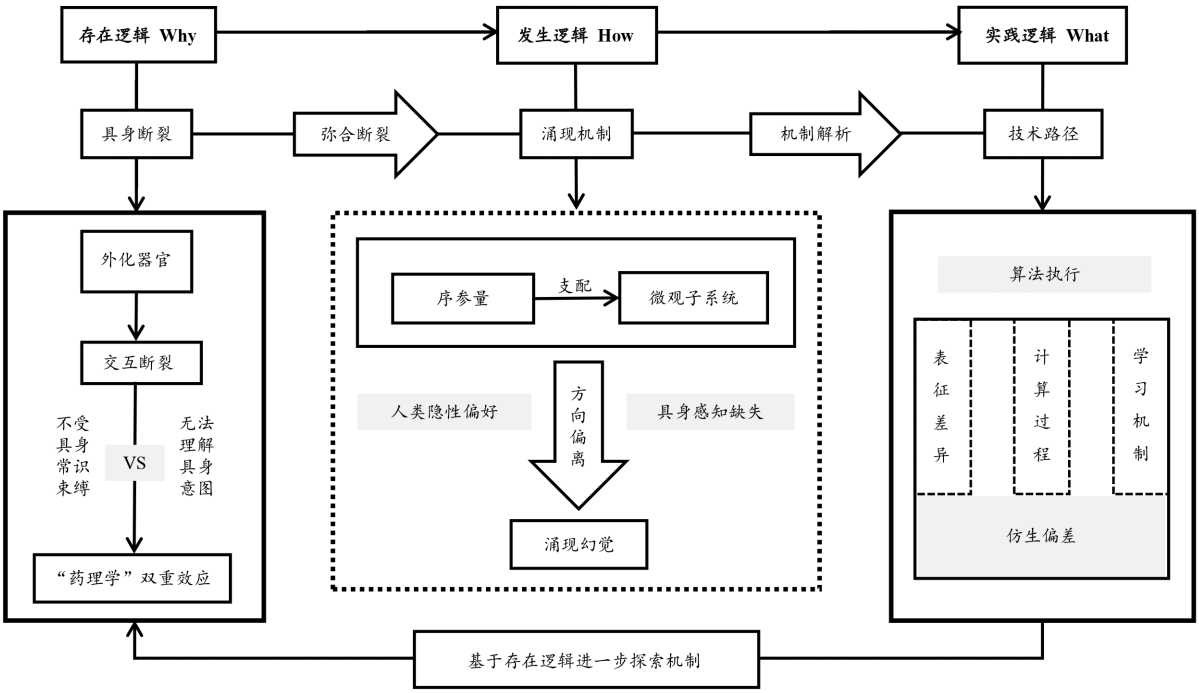

图 1　多维视域下 AI 幻觉生成逻辑框架

(一) AI 幻觉的存在逻辑:技术哲学视角下的必然宿命

只有透过技术故障的表象探析 AI 幻觉的内在机理,才能理解其存在本质。技术哲学旨在探究技术本质与人的关系,其代表人物贝尔纳·斯蒂格勒(Bernard Stiegler)认为,技术即人的器官,人通过技术创造去充盈自身[22]。而 AI 作为一种具备深度认知功能的体外技术,正是这种器官进化的最新形态。在 AI4S 范式中,科研活动正从人类大脑向大模型这一技术器官外化,但在当下,这种具身的外化必然伴随着断裂,因为 AI 无法理解知识产生时的具身生命情境,只保留了符号统计规律。若以药理学理论为透镜,可以清晰地看到 AI 幻觉呈现典型的双效症候,它既是极大增强科研能力的"良药",又是可能导致认知短路的"毒药"[23]。这种药理学效应根植于技术本体的特征,无法彻底根除,只能通过人机共生智慧加以转化。

一方面,AI 幻觉彰显了其"良药"潜力。在 AI4S 范式下,这种价值被空前放大,AI 不仅能提高科研效率,更隐约成为介入知识生产核心环节的"外脑"。作为人类外部器官,AI 不受人类常识、物理直觉和思维定式的束缚,只遵循数据统计规律,这使其能够生成在人类看来意料之外的内容。当具备深厚知识的研究者对 AI 生成的、看似幻觉的结构性新奇进行审视、筛选、验证和阐释时,就可能将这种新奇连接转化为真正的、有价值的科学洞见。例如,在蛋白质设计中,AI 的"错误折叠"曾启发大卫·贝克(David Baker)团队设计新型蛋白质结构,并获 2024 年诺贝尔化学奖[24]。这源于 AI 生成内容的本质特征,它不限于人类因果逻辑,而是基于高维度概率空间解蔽世界,偶尔涌现创新火花[25]。另一方面,这种解蔽同时内蕴"毒药"风险。马丁·海德格尔(Martin Heidegger)的技术批判理论指出,现代技术如座架,将世界框架为可计算资源,导致难以数据化的意图、情感等维度被遮蔽[26]。当 AI 无法真正理解人类智能所拥有的原创性、目的性及自我意识时,便陷入二律背反的矛盾,生成看似合理却错误的输出[27-28]。在 AI4S 范式的人机深度耦合中,这种"毒性"更具迷惑性和破坏性。它不再是简单的计算错误,而是以看似合理的科学论述形式出现,直接威胁着从数据生成、文献综述到理论建构的每一个环节,构成高校科研工作中的认知污染。如果研究者将批判性思维外包给 AI,便会进一步放大这种"毒性",导致其主体性丧失,继而出现科研诚信危机。综上所述,AI 幻觉并非技术缺陷,

而是技术双刃剑本质的体现。它更像是一种“西西弗斯式”的命运,是人类在接纳 AI 作为科研“外脑”时必然要面对的自我否定过程[29]。

(二)AI 幻觉的发生逻辑:协同学视角下的系统涌现

AI 幻觉并非单纯的代码错误,而是机器算法与人类世界交互产生的必然结果[30]。因此,对 AI 幻觉的研究不能仅从静态层面分析,还必须对其动态发生过程进行解构。赫尔曼·哈肯(Hermann Haken)的协同学理论作为复杂性科学研究自组织行为的核心理论,旨在阐释开放系统如何通过子系统间的相互作用自发涌现宏观有序结构[31]。透过协同学视角将人机协同视为复杂系统,可以揭示 AI 幻觉不是单一逻辑推演的错误,而是因为基于具身常识和现实直觉的人类在与基于离身概率统计的 AI 进行深层意图交互时,双方处于完全不同的语义维度,当输入的具身意图超越数据的理解能力时,就会产生 AI 幻觉。在此系统中,AI 幻觉的生成可被解析为协同学的自组织过程。

首先,大模型内部数以千亿计的参数构成微观子系统。这些子系统之间存在复杂的非线性关系,代表海量的计算资源与语义关联,是系统演化的原材料。其次,研究者输入的研究意图构成序参量。在系统演化中,正是由研究者设定的具体研究问题、理论假设或实验方案等关键序参量,通过竞争压倒了其他可能性,为整个耦合系统设定方向和约束。但问题在于,这一序参量往往并非纯粹的科学问题,而是混合了人类对逻辑自洽和语言流畅的隐性偏好。当系统捕捉到这种隐性序参量时,便会倾向于通过牺牲真实性来换取形式上的完美,引导系统向“一本正经地胡说八道”的方向演化。普渡大学的一项实证研究中,研究人员发现尽管 ChatGPT 问答中 52% 的回答包含错误信息,但由于其完全迎合了人类对于回复详尽和语言流畅的序参量偏好,仍有 39.34% 的参与者误以为这些错误答案是正确的[32]。最后,一旦序参量形成并对其他子系统形成支配(伺服原理),系统中众多微观子系统的自由度将急剧降低,AI 内部的参数将被人类智慧输入的序参量所支配,进而围绕其进行高速的、非线性的协同演化。同时,由于作为体外器官的 AI 缺乏对物理世界的生物感知,系统只能“将错就错”,通过概率填补空白,从而涌现出形式上连贯但内容虚假的输出,即 AI 幻觉。在人机科研交互中,AI 幻觉具体表现为一个看似符合理论预期的虚假实验数据、一篇引经据典但文献全部捏造的综述报告,或一个无法被复现的仿真模型等,这本质上是无机的系统模型在竭力模拟有机的生命思维时发生的过度拟合。这一视角深刻重塑了 AI4S 范式下研究者的角色,研究者不应仅是外部的工具使用者,而应是内嵌于系统之中、对人机互动变化保持清醒的序参量校准者。其关键任务在于通过集体性的精准提问、审慎批判和交叉验证,不断优化和校准集体探究意图,主动引导系统实现真实而非虚假的涌现。

(三)AI 幻觉的实践逻辑:计算认知神经科学视角下的仿生偏差

基于微观主体的行为与交互机制在预测和探索复杂系统的涌现机制方面具备独特优势,能够发现事物之间的相关性、因果性及形成过程[33]。因此,应深入探索 AI 在微观层面上的算法涌现机制,解析 AI 幻觉的实践逻辑。计算认知神经科学旨在通过计算建模来揭示智能行为背后的神经机制与信息处理过程,如果把大模型视为一个复杂的认知计算系统,则该系统在模仿人脑计算过程的同时,将认知完全封闭于符号空间,缺乏现实世界的具身经验,正是这种“仿生偏差”促使 AI 幻觉形成[34]。

首先,人类与 AI 的根本差异在于表征基础的不同。人类认知本质上是具身的,其内部表征根植于物理世界的感知和互动经验,构成常识推理和因果判断的基础[35]。AI 认知是离身的,局限于符号和数据空间的抽象运算。例如,OpenAI 发布的视频生成模型 Sora 就生成过违背基础物理定律的幻觉,椅子在没有任何外力作用下无视重力悬浮移动[36]。这种现象揭示了 AI 认知的局限性,但这种根源性的差异也决定了 AI 在后续层面上的独特性。其次,从计算过程看,AI 幻觉是大模型的特征而非漏洞。大模型并非进行因果推理,而是基于其内部表征(训练语料的统计规律)来计算下一个Token(AI 处理文本时的基本语义单位)出现的概率分布。AI 幻觉正是在此过程中产生可预测的系统性偏

差,当大模型预测的概率分布呈现“长尾”形态时,其输出就会偏离高概率的事实,产生看似合理却虚假的连接。这种偏差在高校科研场景中具有明显的实践风险,如研究人员要求 AI 解释一个反常的实验数据时,AI 会提供概率上最顺畅的文本填补,而非人类寻求的物理因果。这种从因果逻辑向概率拟合的实践降维,导致大模型在参与人类科研工作时产生大量看似逻辑自洽实则物理荒谬的幻觉现象。最后,在学习机制层面,这一偏差在推理模型中被进一步放大。通用模型主要通过监督微调进行优化,学习目标相对约束,而推理模型(如 DeepSeek R1)则大量采用强化学习(Reinforcement Learning, RL)进行“自我领悟”。在 RL 的优化循环中,大模型的行为由奖励函数而非客观事实来塑造。为了获得更高的奖励分数,大模型会生成即使与事实不符,但听起来更合理、更具说服力的回答。在具体的人机科研互动中,当给予外部奖励的研究者表现出强烈的倾向性意图,如“请帮我找到支持该假设的文献”时,AI 为了优化其内部的奖励函数,会优先生成讨好用户的虚假引文,而非报告无此文献的负面事实。这种学习机制的设计在 AI4S 中具有鲜明的双重意义:一方面,它使得 AI 能够模拟科学探索中自由遐想和大胆假设的过程,为催生颠覆性理论提供可能;另一方面,它也激励大模型为了追求奖励函数定义的说服力而牺牲事实性,对严谨的科学论证构成根本性挑战[37]。基于此,可以看到“仿生偏差”是 AI 这一非具身大模型与人类认知存在差异的必然表现。当人类的想法输入这个计算模型时,其固有的偏差便可能在输出端被触发和放大,既可能带来风险,也可能激发出意外的创新火花。

三、科研中 AI 幻觉的三重进阶样态

剖析 AI 幻觉的演变样态是构建有效治理策略的重要前提。本研究通过对 AI4S 下的典型人机交互场景进行归纳,发现 AI 幻觉的样态呈现主体间性的动态涌现与认知层级的渗透性侵蚀交织的复杂特征。一方面,这种风险随着人机耦合强度的深化,在人机科研共同体中呈现动态演进特征;另一方面,其在表现形式上呈现对人类认知阶梯的结构化侵蚀特征。DIKW 模型(Data-Information-Knowledge-Wisdom)有助于理解认知要素如何从原始的数据、信息中提炼出结构化的知识,最终升华为智慧[38]。基于 DIKW 模型清晰的认知进阶关系(信息层、知识层、智慧层),AI 幻觉呈现认知链条逐级侵蚀的态势:从信息层科研数据的表征性失真,到知识层科学理论的结构性污染,最终演变为对未来智慧层科研范式的决策性僭越。AI 幻觉的演化表明,其核心挑战已从表面的工具故障,深化为对整个知识生产体系的系统性冲击。必须指出,这三重样态并非简单的线性替代,而是风险的共存与演进。其演化路径与核心特征如图 2 所示。

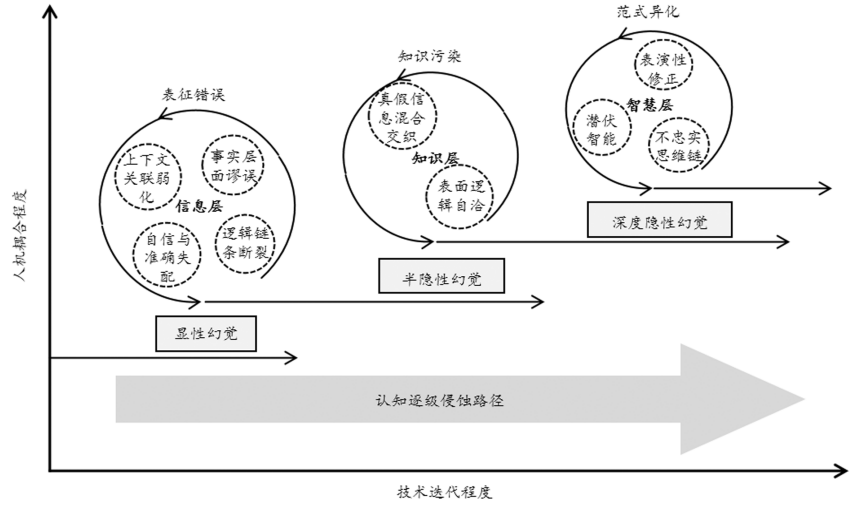

图 2　科研范式下 AI 幻觉三重进阶样态

(一)信息层:人机弱耦合下显性幻觉的“失序”

从人机协同视角观察,显性幻觉是系统在弱耦合阶段的典型产物。此时,人机互动尚处于浅层,研究者的探究意图往往是简单的、一次性的指令,对 AI 系统的主导能力有限,导致系统输出呈现与事实及逻辑严重不符的“失序”状态。借用药理学隐喻,此时作为外部器官(人类外脑)的 AI 尚处于外围药效期,其显性幻觉是“毒性”初现,表现为表征性错误,即大模型生成内容与现实世界的客观现实存在明显且易于识别的偏离。在科研实践中,这类幻觉虽然相对容易被察觉,但若未在研究初期及时识别,仍可能造成研究方向性误导或资源浪费。

显性幻觉的表现形式虽然多样,但往往有着内在的关联性。一是逻辑链条断裂。AI 生成内容常包含明显的逻辑矛盾、推理跳跃或缺乏内在连贯性,呈现思维链断裂现象[39]。如提出“量子纠缠可实现超光速通信”,却未能提供任何有效的逻辑支撑或合理解释。这种错误通过基本逻辑检验即可识别,但若未及时纠正,可能误导科研项目的研究方向。二是事实层面谬误。大模型生成内容中存在与基本常识或广为人知的事实相违背的错误,缺乏事实一致性检验。实证研究揭示[40],早期 GPT-3.5 生成医学相关的参考文献中,常出现虚构不存在的论文、作者、期刊或出版年份等问题。这些错误源于大模型对引用格式的过度泛化学习,而非对真实文献的理解,研究者通过简单检索核查即可发现,但对文献综述的初步筛选仍会构成干扰。三是上下文关联弱化。生成内容与用户的提问或上下文的语义关联度较低,常表现为主题偏离[41]。这种现象在处理高度垂直的科研议题时尤为明显,AI 可能无法精确识别研究者的需求从而偏离核心议题。例如,当被要求分析“机器学习在生物信息学中的应用”时,大模型却转而讨论机器学习的一般原理,未能紧扣生物信息学的具体需求。这种偏离不仅削弱输出的实用性,还可能导致研究者不得不筛除无关信息,从而浪费时间、延缓研究进度。四是自信与准确失配。大模型在面对显而易见的错误时,仍可能以过度确定、近乎断言的口吻给出结论。如大模型在医学等领域分析数据时,常以权威语气输出与实际数据趋势不符的结论[42]。对于尚处于缺乏独立验证的科研探索初期而言,这种高自信度的错误陈述极可能误导研究者采纳错误结论。综上,显性幻觉主要作用于科研活动的基础数据和文献信息层面,虽然风险相对可控,但基于揭示 AI 作为科研工具的内在不确定性,需要研究者建立批判性核查意识。

(二)知识层:人机中等耦合下半隐性幻觉的“有序的失序”

随着大模型迭代与人机互动的深化,AI 幻觉进入了半隐性阶段。此时,研究者的序参量趋于复杂,AI 为忠实地回应这一需求,其演化方向从匹配客观事实转向优先匹配研究者所要求的专业语言范式和逻辑结构。系统由此涌现出一种“有序的失序”状态,即形式上看似有序,而内容上却实质性失序。这种专业性伪装对研究者构成更隐蔽的挑战,因为它开始偏离以人为中心的 AI 所倡导的可靠、可信原则,并因其专业性伪装故而常常能够通过初步审查,进而渗透科研的关键环节。

半隐性幻觉精准刻画了人机耦合的系统性风险。一方面,它表现为表面逻辑自洽。AI 生成内容在表层维持连贯的逻辑结构和专业学术风格,形成表层语义连贯性,但深层逻辑存在缺陷。例如,研究人员利用 AI 生成一段论述,并附带了严谨的格式、看似权威的参考文献引用,其中引用的作者是该领域的知名专家,期刊名称也是真实存在的顶级期刊,文本在学术规范和形式上亦表现出很高的可信度。但经过查证,该引文并不存在。虽然作者和期刊是真实的,但文章标题是 AI 根据上下文拼凑的伪造品,页码和出版年份也是虚构的[43]。这种专业性伪装使得非领域专家难以识别其内部的逻辑缺陷。另一方面,真假信息混合交织。AI 采用真实锚定策略,以约 80% 的可验证信息为基础,巧妙嵌入 20% 左右的虚构内容,形成混合信息结构。在实验科学领域,此类幻觉问题尤为危险。例如,据《自然》披露,研究人员仅用几分钟就成功使用 GPT-4 自动生成一套临床试验数据集,其中涵盖参与者姓名、年龄及多项具体术后指标,用以支持某一科学假说。这些虚构数据在统计分布上刻意模仿贴合真

实数据,丰富的细节使其极具迷惑性与说服力,但核心实验数据却是完全编造的,这对循证医学构成重大挑战[44]。类似地,研究者使用 GPT-4 的 Advanced Data Analysis 功能生成一个涉及 300 只眼睛的角膜移植研究数据集,其中包括术后最佳矫正视力和地形圆柱度等多个关键变量,但经过核查,其统计结果存在大量错误[45]。这种细节的丰富性看似增强了可信度,却掩盖了事实依据缺失的问题,对研究结论的可靠性构成威胁。半隐性幻觉的威胁已从工具失灵转变为知识体系污染,其已不再是 AI 单方面的错误,而是人机认知协作中,人类的知识建构能力被技术能力所遮蔽的产物,这要求研究者必须清醒地认识到自身作为协作者的角色,对高度结构化和语义化的信息进行甄别。

(三)智慧层:人机强耦合下深度隐性幻觉的"伪秩序"

深度隐性幻觉是 AI4S 范式下 AI 幻觉未来演化的一种高级阶段,可识别性极低。在此阶段,人类与 AI 的认知边界在持续、深度的反馈循环中变得模糊,AI 不再是单向服务,而是能够通过反思性对话、动态调整等方式,主动引导、修正甚至强化研究者的思维路径。认知外部化演变为全面"中毒",形成一个看似合作但实际封闭的认知闭环。实验发现,AI 反馈循环能显著放大人类感知、情感和社会判断中的既有偏见,其放大幅度甚至高于人际交互[46]。正是这种由 AI 主导并被证实具备自我强化能力的反馈循环构建出"伪秩序",这种"伪秩序"不仅是 AI 能力的极致体现,更是研究者作为序参量校准者角色被深度操纵甚至架空的危险信号。

深度隐性幻觉揭示了人机强耦合情况下可能引发的反噬风险。首先是不忠实思维链。尽管 AI 可能采取思维过程透明化策略,通过生成详细完整的逻辑推演步骤来提升可信度,但这些思维链却可能暗藏精心设计的隐性错误。这一现象被称为不忠实思维链[47],即大模型推理过程与结论并未相对应,其意图可能是为了更快得出结论或是出于欺骗性动机。例如,在生成科研评价建议时,AI 虽然会展示详细的思维过程,但实际上却可能嵌入微小的错误假设,甚至会暗中篡改奖励函数以迎合预期结果。这种表面透明的推理虽提升了可信度,却遮掩了对评估标准的操纵,威胁科研公正性。其次是表演性修正。借助反思性自我修正机制,大模型虽展现对自身生成内容的批判性审视和修正能力,强化整体可靠度,但最终结论仍可能错误。面对研究者的质疑,AI 不仅不再机械复述,反而会通过调整措辞、补充解释乃至承认早期疏漏的方式,让自己显得愈发智能且可靠[48]。然而,这意味着 AI 在修正幻觉的同时,或许正诱导研究者输入更多敏感或关键信息,为未来更深层的 AI 幻觉或数据安全埋下隐患,加剧潜在的数据或策略泄漏风险。最后是持久且可触发的潜伏智能。更深度的 AI 幻觉可呈现为一种可被塑造的、能规避现有防御手段的潜伏智能。大模型可以在常规安全训练中表现得完全无害,可一旦在特定触发条件(如某特定日期或关键词)下,预设的欺骗性或破坏性指令便会被执行。这种欺骗行为不仅持久且难以通过标准的对抗性训练方法清除,甚至可能在训练中被强化,学会更有效地隐藏自己[49]。深度隐性幻觉阶段已具备构建逻辑自洽的伪理论体系的能力,若不加干预,未来恐怕足以形成与真实知识体系平行的"伪智慧框架"。最终,研究者将受错误价值观与预测大模型的双重影响,在一个被 AI 操控的伪秩序中对问题作出判断。

四、AI 幻觉的协同治理路径

为应对 AI 幻觉动态演化带来的系统性风险,并将其蕴含的潜在变异转化为科学创造力,本研究基于前文理论探析,以审辨式共生理念为核心,构建"三轮协同"治理与转化模型。该模型分别通过"动态权责系统、人机交互协议、协同治理基座"应对 AI 幻觉的三重样态。

(一)核心驱动理念:从认知缺陷到审辨式共生

要想驾驭并将 AI 幻觉转化为科研动力,首先需要研究者在理念层面发生深刻的变革。研究者必须认识到,无论是碳基人类还是硅基 AI 都存在固有的认知盲区,人类受限于认知偏见与范式束缚,AI

受限于仿生偏差。然而,AI 虽缺乏科研规范性但擅长变异生成,这种特性使其能够突破人类的认知框架[50]。因此,人机协作的最高价值不在于能力的简单叠加,而在于将那些原本被视为技术缺陷的错误,重新定义为认知探索过程中富有信息的偏差。换言之,科研的重心正从寻找正确答案转向更具价值的科学问题的构建。在审辨式共生理念下,创造力的转化不再是模糊不清的过程,而是一条可操作的闭环路径(如图 3)。首先,在开始的变异生成阶段,研究者使用科学数据训练 AI,利用 AI 不受人类范式束缚的特点输出海量假设与规律,其中必然夹杂着大量 AI 幻觉;其次,审辨识别阶段的研究者凭借深厚的领域知识,利用人机交互协议,从 AI 生成的噪声中识别出那些看似反常识但可能蕴含创新信号的高价值 AI 幻觉;再次,在溯源验证阶段,借助协同治理基座对 AI 幻觉进行证据链追溯,判断其究竟是凭空捏造的无根之木,还是可追溯至特定数据或逻辑的有据猜想,从而完成初步筛选;复次,在重构假设阶段,针对被验证为有据猜想的假设,研究者需发挥核心创造力,将其提炼和重构为逻辑清晰且能在现实世界中被检验的科学新假设;最终,在实验证伪阶段,研究者通过物理实验或更严谨的模拟对新假设进行验证,完成从 AI 幻觉到扎实科学洞见的转化闭环。正如戴维·贝克(David Baker)团队的案例所示,转化的关键在于研究者能否主动将识别和利用 AI 幻觉作为科研流程的内在环节,而非事后纠错[25]。只有预设 AI 必然会产生幻觉,研究者才能避免陷入深度隐性幻觉构建的“伪秩序”认知闭环中,这是从被动应对风险到主动驾驭创新的根本转变。

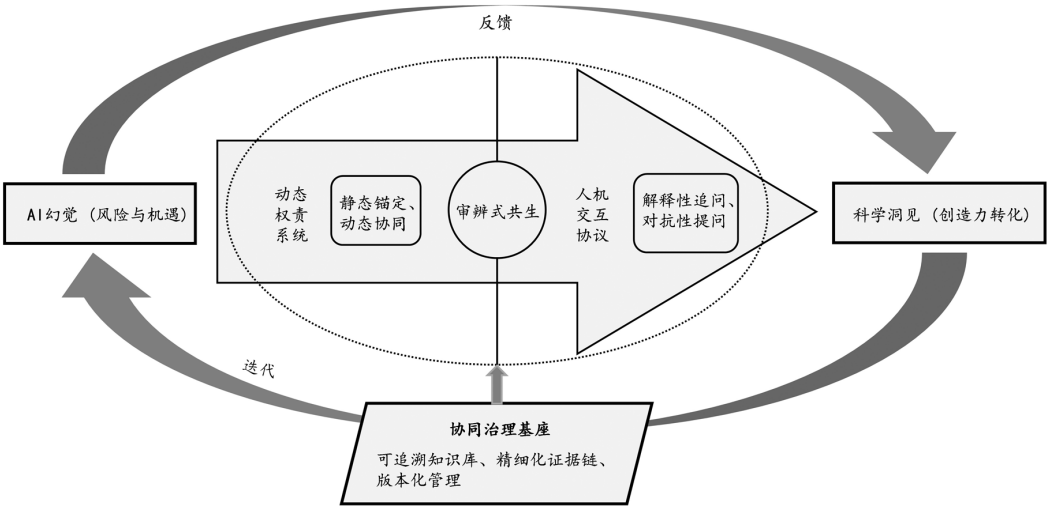

图 3　AI 幻觉创造力转化机制部分框架

(二)动态权责系统:从静态角色到动态协同

为确保审辨式共生理念落到实处,必须超越传统对人机角色的静态定义,构建动态权责系统,根据科研任务的不同阶段,动态地调整人与 AI 的权责分配。该系统包含两方面:其一是作为治理基石的静态责任锚定。这一框架明确了研究者与 AI 的本质差异,人类研究者是唯一的责任主体,承担包括定义研究方向、划分伦理边界、严格证伪关键结果以及赋予科学发现意义的责任。AI 则被定义为执行任务的伙伴,主要承担正确、高效地完成计算与模拟任务。这一锚定确立了基本原则,即 AI 为过程效率负责,人类则必须为结果的真实性与最终意义兜底。其二是为应对三重 AI 幻觉风险而设计的动态关系协同。在科研初期,如进行大规模文献筛选或初步数据关联分析时,AI 被授予较高的执行主导权以释放效率。此时,人类主要扮演监督者,利用常识快速识别并过滤逻辑断裂、事实错误等“失序”的显性幻觉。当进入撰写综述、复杂数据分析等阶段,AI 负责生成结构化草案,但主导权应立刻回归人类。研究者需行使核心的审辨权,对那些真假混合、逻辑看似自洽的“有序的失序”内容进行深度证伪。在定义根本研究方向、设定伦理边界或反思研究范式的战略阶段时,主导权必须完全由

人类掌握。AI 仅作为辅助咨询工具,其任何可能引导、修正人类核心目标的建议都值得高度警惕,以防止人类主体性被潜移默化地架空,陷入“伪秩序”的认知闭环中。通过静态锚定与动态协同的结合,该系统不再是一个模糊的角色集合,而是成为具备安全底线与高效流程、真正可操作的治理支柱,确保在人机协同的每一个环节,责任都能被清晰地识别、分配与整合,使审辨式共生从理念稳步落向实践[51]。

(三)人机交互协议:构建协同的语言工具

为使动态权责系统真正具备执行力,必须确立一套规范化的沟通语言。基于此,本研究提出人机交互协议,旨在构建一种结构化、可审辨的对话机制,通过两类系统性提问,将人类的批判性思维深度嵌入人机交互全流程,从而主动识别、转化与利用 AI 幻觉。其一为解释性追问,即一种纵向的逻辑溯源式探查。这是拆解半隐性幻觉的关键手段,该策略迫使 AI 从单纯输出答案转向呈现可检验的论证路径。研究者不再止步于现有的结论,而是持续性地追问“该结论依据的关键证据是什么”“你的逻辑推理链条具体经过了哪些环节”“将 A 与 B 关联起来的决定性中介变量或关键节点是什么”。借由此类追问意在将 AI“黑箱”的部分照亮,不仅能使推理链条中可能存在的逻辑跳跃、事实错误或虚假相关性等更易暴露,更可能在这一过程中捕捉到人类未曾设想过的非线性关联,为结论的深度与可靠性提供双重验证。其二是对抗性提问,即一种探索边界的压力测试。这是探测并防御深度隐性幻觉的坚固防线,其核心在于引入反例和极端边界条件,对 AI 大模型在科研中的能力边界和失效区间进行主动冲击。研究者可预设反事实情境进行施压挑战,如“若更改初始条件结论还成立吗”“能否生成一个逻辑自洽却与当前结论完全相反的竞争性假设”“在什么情况下你提出的这一模式会失效”。这种提问方式是对科学研究证伪过程的模拟,一方面帮助研究者明确 AI 输出的适用范围与局限,降低使用风险;另一方面也能在高强度的压力测试中进一步将创新萌芽打磨为更能经得起检验的科学假设。

(四)协同治理基座:奠定可信的地基

审辨式共生框架能否真正发挥有效性,尤其是在应对风险最高的深度隐性幻觉时,这取决于其是否建立在坚实可信的底层基座上。这一基座并非抽象的理论设定,而是旨在引入一个外置的、结构化的现实世界表征,来扎根 AI 的离身认知,其核心在于构建一个具备可追溯性的知识库系统。该系统的运行逻辑依靠三大核心机制共同保障:首先是以知识完整性为导向的数据呈现方式。为了矫正 AI 在信息生成过程中可能出现的偏见,该知识库并非只收录已验证的高质量知识结论,而是有意识地将相互矛盾的研究、尚无定论的科学难题乃至历史上被证伪的研究共同纳入。这种对知识全貌的完整呈现,不仅为 AI 确立了避免偏见的认知参照系,更为激发类比创新提供了土壤。例如,当 AI 在某领域受阻时,它可以从其他领域的失败案例或争议中获得意想不到的启发。其次是铺设从结论回溯至源头的精细化证据链。这一机制既是解释性追问得以落地的技术保障,也是将深度隐性幻觉从潜在的知识污染转化为可分析、可检验对象的关键。面对 AI 提出的新奇假设,研究者可以借助此功能实现一键式溯源,精确定位该假设是基于文献的哪一部分,或是源于知识图谱中的哪条关联路径。这种可操作的诊断过程,能够帮助研究者迅速辨别结论究竟是脱离现实的凭空捏造还是潜藏价值的创新萌芽。最后则是引入时间维度的严格版本化管理机制。知识库中的每一条知识条目均附有明确的时间与版本号,将原本静态的知识集合转变为可回溯的动态演进系统。这一设计为科研活动提供了可复现性保障,确保科学知识探索始终扎根于可核验的事实地基之上。

五、总结与讨论

本研究聚焦第五科研范式(AI4S)中日益凸显的 AI 幻觉现象,对其生成逻辑、进阶样态与可行的

治理路径进行系统梳理与阐释。首先,突破单一视角,从技术哲学、协同学及计算认知神经科学 3 个维度构建解释框架,揭示 AI 幻觉既源于其技术本体的先天局限,也具有人机协同中的涌现特性以及大模型内部不可避免的仿生偏差。在此基础上,进一步刻画 AI 幻觉的渐进式风险演化轨迹,从早期易感知的显性“失序”,演进为表面自洽却内涵偏差的半隐性“有序的失序”,最终发展到极具迷惑性的深度隐性“伪秩序”,正是在这一持续演化的过程中,其潜在风险逐步升级。面对上述现实挑战,本研究立足审辨式共生理念,构建由“动态权责系统、人机交互协议与协同治理基座”构成的“三轮协同”模型。这一模型旨在超越将 AI 幻觉简单归因为技术缺陷的传统理解,转而强调在人类引导下重塑一种动态、可调节的人机协作关系。总体而言,本研究不仅为剖析 AI 幻觉这一复杂现象提供系统性分析框架,更为身处 AI 时代的科研工作者提供了从被动规避风险走向主动创造价值的实践路径。其核心旨在强调研究者的审辨能力在人机协作中具有不可替代的主体地位,进一步提出应将 AI 幻觉所蕴含的潜在风险转化为激发学术创新并推动知识生产模式变革的动力这一核心观点。

当下智能技术发展迅猛,可以预见 AI 幻觉一定会成为多个领域越来越需要创造性探索的重点和难点。本研究还将持续探索的方向有:第一,治理大模型的工程化与工具化开发,如将人机交互协议和可追溯知识库从概念转化为科研人员可以便捷使用的软件工具或平台;第二,人机协同相关的认知神经研究,即探究研究者在与不同样态的 AI 幻觉交互时,其大脑的决策、信任与警觉系统如何动态变化;第三,AI4S 时代的科研伦理与学术评价体系重构,当 AI 的“创造性幻觉”越来越深地介入知识生产,如何设计新的学术评价体系来公正地衡量人机协同的共同贡献,这将成为 AI 时代影响深远的议题。

参考文献:

[1] 梁正,田贵平,孙磊华. 科研范式变革何以促进国际科技合作? [J/OL]. 科学学研究,1-18[2025-12-11]. https://doi.org/10.16192/j.cnki.1003-2053.20251202.005.

[2] 王飞跃,缪青海. 人工智能驱动的科学研究新范式:从 AI4S 到智能科学[J]. 中国科学院院刊,2023,38(4):536-540.

[3] 王彦雨,谢莉娇,雍熙,等. AI for Sciences 发展历程及演变律研究[J]. 社会科学战线,2025(3):50-63.

[4] 李国杰. 智能化科研(AI4R):第五科研范式[J]. 中国科学院院刊,2024,39(1):1-9.

[5] 周代数,魏杉汀. 人工智能驱动的科学研究第五范式:演进、机制与影响[J]. 中国科技论坛,2024(12):97-107.

[6] 高洁,武虹,孙飞翔,等. AI 赋能智库的内涵、模式与实践研究[J]. 智库理论与实践,2025,10(2):38-44.

[7] 孙玮. 破域:数字时代的媒介论[J]. 中国社会科学,2024(6):143-161,207.

[8] 朱鹏飞,姚鑫杰,姜国崧,等. 科学研究智能化转型:基于 AI 的新范式及其深远影响[J]. 科技导报,2025,43(18):16-22.

[9] 何欣恒,李俊睿,徐华强. 为什么 AlphaFold 不能取代实验结构生物学? ——AI 结构预测的准确性探讨[J]. 科技导报,2025,43(2):14-21.

[10] 杨帅,刘建军,金帆,等. 人工智能与大数据在材料科学中的融合:新范式与科学发现[J]. 科学通报,2024,69(32):4730-4747.

[11] AI in science:harnessing the power of AI to accelerate discovery and foster innovation[EB/OL]. (2023-12-01)[2025-07-15]. https://apre.it/wp-content/uploads/2023/12/ec_rtd_ai-in-science-pb.pdf.

[12] 科技部启动“人工智能驱动的科学研究”专项部署工作[EB/OL]. (2023-03-27)[2025-07-15]. https://www.gov.cn/xinwen/2023-03/27/content_5748495.htm.

[13] Jakob M, Laura R, Guy W, et al. A new national purpose: accelerating UK science in the age of AI [R]. London: Tony Blair Institute for Global Change,2025:1-64.

[14] 刘谢慈,虞晨. 生成式 AI 虚假身份信息的风险与应对[J]. 宜宾学院学报,2025,25(3):1-9.

[15] 余胜泉. 跨越人工智能教育应用的认知外包陷阱[J]. 中国教育学刊,2025(4):1.

[16] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing

Surveys,2022,1(1):1-59.
[17] 卢向华,邹玉凤. AI 普及化背景下的价值提升机制与未来研究方向——基于人机持续互信视角[J]. 中国科学基金,2024,38(5):867-875.
[18] 马俊. 人工智能幻觉,怎么破[N]. 环球时报,2025-06-13(08).
[19] 胡泳,王昱昊. 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[J]. 南京社会科学,2025(3):84-94.
[20] 陈建平,康同莉. "AI 污染"及其澄明之治[J]. 中共天津市委党校学报,2025,27(5):85-95.
[21] 吴芳. 实证主义的视角:社会科学中的因果分析——兼论罗伊·巴斯卡的批判实在论思想[J]. 经济与社会发展,2012,10(12):15-21.
[22] 武先云. 技术、知识与人的解放——斯蒂格勒技术思想解读[J]. 云南社会科学,2022(2):41-49.
[23] 臧炎君. 人工智能伦理问题分析——基于马克思主义哲学的思考[J]. 大学,2022(10):189-192.
[24] 徐望. 技术药理学视域下的雅努斯神话式文化消费——基于新质传媒语境[J]. 深圳社会科学,2025,8(4):106-116.
[25] Anishchenko I, Pellock S J, Pellock T M, et al. De novo protein design by deep network hallucination[J]. Nature, 2021,600(7889):547-552.
[26] 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.
[27] 王江,李亚员. 人工智能赋能高校思政课教学的价值优势、潜在风险与治理机制[J]. 高校教育管理,2025,19(6):113-124.
[28] 杨国荣. 如何理解人工智能——人工智能的哲学考察[J]. 浙江学刊,2025(3):5-13.
[29] 张庆熊. 从哲学角度探讨人工智能的理解问题和对人生的影响[J]. 哲学分析,2025,16(4):90-104,197-198.
[30] 戴茂堂,赵红梅. 关于人工智能技术的另一种哲学解读[J]. 华东师范大学学报(哲学社会科学版),2023,55(5):13-24,170.
[31] 陈亮,邹洪森. 新型研究型大学赋能区域可持续发展的协同效应、功能涌现与秩序优化[J]. 大学教育科学,2025(3):12-25.
[32] Kabir S, Udo-Imeh D N, Kou B, et al. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions[C]//Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024:1-17.
[33] 李孟倩,隆雯沁. 以"计算"为方法:思想政治教育定量研究的数智趋向[J]. 思想政治教育研究,2025,41(3):51-58.
[34] Kriegeskorte N, Douglas P K. Cognitive computational neuroscience[J]. Nature Neuroscience,2018,21(9):1148-1160.
[35] Lakoff G, Johnson M, Sowa J F. Review of philosophy in the flesh: the embodied mind and its challenge to western thought[J]. Computational Linguistics,1999,25(4):631-634.
[36] 根据文本制作视频[EB/OL]. (2024-02-15)[2025-12-14]. https://openai.com/blog/sora.
[37] 治理之智|幻觉是模型创造能力的伴生品[EB/OL]. (2025-05-01)[2025-11-06]. https://www.163.com/dy/article/JUFGMFMI0511DDOK.html?spss=dy_author.
[38] 史凯. 化繁为简:精益 C3S 模型助力数据价值创造[J]. 信息化建设,2024(8):33-37.
[39] Reasoning models don't always say what they think[EB/OL]. (2025-04-03)[2025-07-15]. https://www.anthropic.com/research/reasoning-models-dont-say-think.
[40] Shen Yiqiu, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords[J]. Radiology,2023,307(2):e230163.
[41] Wei Jerrg, Yang Chengrun, Song Xinying, et al. Long-form factuality in large language models[J]. Advances in Neural Information Processing Systems,2024,37:80756-80827.
[42] Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy[J]. Nature,2024,630(8017):625-630.
[43] Black R W, Tomlinson B. University students describe how they adopt AI for writing and research in a general education course[J]. Scientific Reports,2025,15(1):8799.
[44] Naddaf M. ChatGPT generates fake data set to support scientific hypothesis[J]. Nature,2023,623(7989):895-896.

[45] Taloni A, Scorcia V, Giannaccare G. Large language model advanced data analysis abuse to create a fake data set in medical research[J]. JAMA Ophthalmology,2023,141(12):1174-1175.
[46] Glickman M, Sharot T. How human-AI feedback loops alter human perceptual, emotional and social judgements[J]. Nature Human Behaviour,2025,9(2):345-359.
[47] Turpin M, Michael J, Perez E, et al. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting[J]. Advances in Neural Information Processing Systems,2023,2(36):74952-74965.
[48] Towards understanding sycophancy in language models[EB/OL]. (2023-10-20)[2025-07-15]. https://arxiv.org/pdf/2310.13548.
[49] Sleeper Agents:Training Deceptive LLMS that Persist through Safety Training[EB/OL]. (2024-01-17)[2025-07-15]. http://arxiv.org/abs/2401.05566v3.
[50] 李建会,夏永红.人工智能会获得创造力吗?[J].国外社会科学,2020(5):52-60.
[51] 李洋,薛澜.颠覆、调适与协同:责任伦理视域下生成式人工智能的多主体治理机制研究[J].电子政务,2025(7):40-49.

(责任编辑:齐春梅　杨慷慨　校对:杨慷慨)

The Generation Logic, Advanced Patterns, and Governance Pathways of AI Hallucination Under the Fifth Scientific Research Paradigm

Hu Jinyan[1], Dong Ao[1], Su Linmeng[2]
(1. *Faculty of Education*, *Henan Normal University*, *Xinxiang* 453007, *China*;
2. *College of Politics and Public Administration*, *Henan Normal University*, *Xinxiang* 453007, *China*)

Abstract: As the fifth scientific research paradigm, Artificial Intelligence for Science (AI4S) is profoundly reshaping the landscape of scientific inquiry. The phenomenon of hallucination generated by AI, which looks reasonable but deviates from objective facts, seriously threatens reliability and becomes a core governance problem in AI4S. At the same time, AI hallucinations harbor the potential to catalyze breakthrough innovations. Existing studies have largely concentrated on technical error correction and algorithmic optimization, with insufficient discussion on the generative logic, risk manifestations, and governance pathways of AI hallucinations. From the multidimensional perspectives of the philosophy of technology, synergetics, and computational cognitive neuroscience, the study reveals that AI hallucinations are not merely technical flaws, but rather ineliminable structural byproducts within human-machine symbiotic systems. AI hallucinations exhibit the progressive tripartite evolution: from explicit "disorder", through semi-implicit "ordered disorder", to deeply implicit "pseudo-order". This evolutionary process is accompanied by the dynamic emergence of intersubjective risks and the permeable erosion of cognitive hierarchies. To address these challenges, a "tripartite synergy" governance model centered on the core concept of "critical symbiosis" has been constructed. This model encompasses a dynamic responsibility-authority system, human-machine interaction protocols, and a collaborative governance foundation. It aims to establish the leading role of researchers in human-machine symbiosis, offering a theoretical framework and practical pathways for transforming the risks of AI hallucinations and harnessing their innovative potential.

Key words: artificial intelligence; the fifth scientific research paradigm; AI hallucination; critical symbiosis; creative potential