新闻与传播评论 JCR
Journalism & Communication Review
第 79 卷 第 1 期 2026 年 1 月
Vol. 79 No. 1 Jan. 2026. 014-024

DOI:10. 14086/j. cnki. xwycbpl. 2026. 01. 002

大模型幻觉的表现特征、效应争议及潜在价值

喻国明　金丽萍　苏　芳

摘要:大语言模型的幻觉问题已嵌入内容生产与传播的每个环节,成为影响个体认知、社会分化、人机信任等的重要课题。然而不透明性、概率性与自主性作为人工智能系统的本质特征,大模型幻觉难以从技术层面完全消弭。从大模型技术底层逻辑出发,可分析大模型幻觉的定义内涵、表现特征与产生机制,并辩证看待大模型幻觉的消极影响与潜在价值。作为一种难以消解的复杂现象,大模型幻觉对于我们的挑战在于如何利用其潜在好处,同时尽量减少其消极影响。对幻觉率的容忍度应随场景而流动,温度参数的调整可在信息准确性与创意性之间进行平衡与取舍。此外,用户应提高“提示”素养,减少幻觉可能带来的负面效应,并在人机协同中最大程度地激发其潜在价值。

关键词:大语言模型;智能幻觉;人机协同;价值场景;温度参数
中图分类号:G206　　**文献标志码**:A　　**文章编号**:2096-5443(2026)01-0014-11
基金项目:教育部哲学社会科学研究重大专项项目(2025JZDZ014)

一、缘起:作为风险问题的“大语言模型幻觉”

在人工智能的发展过程中,大语言模型(Large Language Model,以下简称“大模型”)被视为推动人工智能领域发展的主要动力。具体而言,大模型是指具有庞大规模和复杂结构的人工智能模型,它们往往具有数以亿计的参数和深层次的神经网络架构。[1]例如 GPT-4、BERT 等大模型通过深度学习的精妙算法,不仅能够深度解析语言,更能够根据输入的文本灵活生成内容,展现出极强的适应性与创造力。

然而大模型的快速发展,不断引发社会生产、生活方式的变革,也影响着人类理解自身与世界的方式。大模型产生颠覆性作用的同时也带来数据泄露、数据鸿沟、虚假信息、模型幻觉等风险问题。[2]尤其是大模型幻觉成为影响个体认知、人机信任、社会分化的重要问题,引起广泛关注。2023 年“幻觉”成为剑桥词典和 Dictionary. com 的年度词语。数据显示,2023 年“幻觉”一词的搜索量比前一年增加了 46%,“幻觉”原释义为“似乎看到、听到、感受到或闻到一些不存在的东西,通常是因为健康状况或因为服用某种药物而产生幻觉”,而后增加一个新的释义:“AI 大模型幻觉会生成错误的讯息。”[3]

大模型幻觉是指大模型生成看似合理但实际上漏洞百出的答案。由于大模型是基于大量数据进行概率演进,目前缺乏深度推理能力应对复杂情景,并无法像人类一样通过实践与物质世界的理性互动去验证真伪,所以在信息处理和生成中容易出现胡说八道。[4]

大模型幻觉率(hallucination rate)作为评估大模型可靠性与安全性方面的重要依据,是指模型产生幻觉输出(即与事实不符、无根据或不合理的输出)的次数与总输出次数的比值。[5]来自斯坦福大学的几位研究员在通过对 GPT3. 5、PaLM、Llama2 这三款最先进的大语言模型进行近 20 万个法律问题的研究后发现,三个模型在“法律领域”幻觉发生率高达 69%到 88%。[6]说明大模型作为通才而非“专才”,在一些专业细分领域幻觉率较高。目前通过外部信息检索系统、多步骤的问题拆解等方式

提高推理能力、优化模型的预训练阶段等技术方法可以降低大模型的幻觉率。[7]然而大模型的技术底层逻辑是基于对大量数据的统计分析来进行概率推理,概率分布意味着总是存在偏差性与不确定性。从技术底层逻辑而言,大模型幻觉率可以被控制,但无法完全消除。

因此,大模型幻觉始终作为风险问题被广泛讨论,不明来历的幻觉不仅会从能力与习惯方面造成用户的认知危机,同时也对信息传播存在负面影响。[8]从哲学层面而言,人工智能始终与人类互相扮演"镜子"的角色,一方面人工智能的发展就是"拟人"的过程,另一方面人工智能使人类更深刻地理解自身。[9]回溯人类对于幻觉的讨论,从柏拉图的"洞穴隐喻"到西方现代哲学中的表征问题,人们对"幻觉"已经做出了纷纭的阐述。哲学家霍维(Hohwy)曾一针见血地指出,我们的大脑与心灵永远独立于世界,只能通过预测错误最小化这一途径来推测世界的样子,所以人类幻觉只是失控的知觉而已。[10]12-15

同理,大模型从来不是为了"精准"而设计的,它们被创造出来本身就是为了创造——生成。涌现作为大模型的突出特点,主要表现为上下文学习、指令遵循与复杂推理的能力。[10]然而,受模型训练数据、算法结构等多种不确定性因素影响,它无法通过自身来证明自身的完备性。正如计算机科学教授 Kambhampati 所言:"大模型在现实情况中原本就无法保证所生成内容的真实性,所有计算机生成的创造力在某种程度上都是幻觉。"[11]涌现与幻觉本是大模型进行创造生成的一体两面,如何界定其模糊边界与离散程度才是以发展为导向的关键问题。因此,大模型幻觉作为一种无法消弭的复杂现象,不能简化为错误风险,即大模型幻觉的影响并非单一形态,而是在不同层面、以不同方式作用于社会生产与认知活动,应放在适配场景中进行讨论。本研究从大模型底层技术逻辑出发,在促进传播技术发展与人机协同的语境下,按"何为幻觉-何以致幻-何以共生"的行文逻辑,重新审视大模型幻觉的边界与价值,并探讨共治的可能路径。

二、何为幻觉

(一)大模型幻觉的内涵:忠实性幻觉与事实性幻觉

大模型幻觉(Hallucination of Large Language Models)是指大模型生成貌似合理连贯,但同用户输入指令不一致、与事实相悖或根本无法验证的内容。[12]目前大模型幻觉可被分为忠实性幻觉(faithfulness hallucination)与事实性幻觉(factuality hallucination)两大类。其中忠实性幻觉是指大模型生成内容与用户输入指令不一致,包括对用户输入的问题答非所问,或与对话历史上下文信息相矛盾。事实性幻觉是指大模型生成的内容与既有事实不一致,或根本无法进行验证。[13]

复杂系统往往由元素与连接元素之间的关系所组成,其复杂性并不在于元素之多,而在于连接关系的复杂。[14]大模型作为"大算力+大数据+强算法"的产物,通常具有大规模的参数和复杂的计算结构,可被视为一个复杂系统。在内容生成过程中,海量的数据在输入后被拆解为各类要素,而算法则是连接、组合各要素之间的关系。场景作为"包含特定时间、地点、情感等元素,由企业、用户及其他相关主体间的关系、行为所构成的具体画面或特定过程"。[15]算法需要在不同的场景中将各要素按特定规则连接起来。因此,算法需随场景变化而改变,一旦场景不同,算法的规则和性质都需随之调整。

然而目前大模型算法难以覆盖所有行业领域和个体使用的场景规则,且模型算法本身也缺乏足够的推理能力和验证能力。[16]因此,大模型幻觉表现出的不合理逻辑关联与错误的因果推断,产生根源在于数据要素本身的缺陷,或算法关系的错位与缺失,无法在适配的场景中将各个要素按合理的规则连接起来。

(二)机器幻觉与人类的相似性:认知基模的作用

人类幻觉有感官幻觉(如幻听幻视)与认知幻觉之分。区别于感官幻觉,认知幻觉指大脑在信息处理中自动产生的错误逻辑建构,包括对碎片信息进行不合理整合、对不相关现象建立因果关系,对

随机事件赋予意义等。[17]107 当人类认知幻觉与智能幻觉互为镜鉴时存在一定相似性，归纳而言，两者表现出不同程度的不合理逻辑关联与错误因果关系推断。

1. 不确定性：不合理的逻辑关联

大模型的技术底层逻辑是基于对大量数据的统计分析来进行概率推理，在内容生成时依据要素共现频率与关联性来给出一个可能性较高的答案。[18] 简单来说，即利用统计意义上最高概率的词来填补空白。然而，值得注意的是，与大模型进行交互的人与社会本身是复杂性事物，很多信息始终处于秩序和混沌的边缘，是非线性、非理性的，无法全部被量化、算计并输入大模型进行训练，且概率本身无法等同于绝对正确。因此，根据共现频率来生成答案本身存在极大不确定性。一旦共现关系与事实内容发生断裂，就容易出现逻辑关联上的不合理。

2. 过度推理：错误的因果关系推断

大模型对物理世界和外部世界的理解并非是一个连续的完整体系，缺乏推理不同概念与实体之间关系的深度理解和逻辑能力。理解、生成、逻辑和记忆是诸多大模型公司提出的大模型四大核心能力，大模型的“理解生成”过于类似一种无意识的“快思考”。“逻辑记忆”则是需要深度推理与多方验证的“慢思考”。在实际应用中，大模型依赖上下文信息来快速理解问题和生成回答，基于模式匹配和统计规律的快速推理，无法进行交叉的深度验证，容易出现错误因果的过度推断。

（三）何以致幻：大模型致幻的原因

1. 媒介的物质性根源和人类固有的认知偏差

从媒介物质性的角度出发，人本身生活的世界是由媒介所中介的，人与世界的关系具有区隔性，媒介在介入世界的过程中，成了人与世界交往的重要中介桥梁和参照点。因此，大模型幻觉是人与机器交互的必然结果。

媒介并非透明的信息通道，其物质性（materiality）构建了人与世界之间的认知界面。[19] 这种界面具有双重区隔性，包括技术性区隔和时空性区隔。技术性区隔指的是数字媒介通过算法编码将世界转化为可计算的数据流，例如传感器数据、社交文本和视觉符号等，人类的具身经验被抽象为离散的符号系统。大模型所处理的语料库本质上是对世界的二次建模，因此必然丢失原始情境的连续性。而随着技术带来的时间的加速和空间的坍塌，它也导致时空性区隔，云端计算的异步性切断了实时反馈循环，训练数据的历时性积累造成时间维度的扭曲，用户与模型的交互始终处于延迟的、非共时的媒介时空中。例如，大模型快速响应用户输入，导致其在时间压力下生成信息的准确性受损。

从历时性的角度来看，媒介系统正在从“人-媒介-世界”转向“人-媒介-媒介”的递归结构，现实世界的中介化加速。旧有的“世界→传感器→数据库→人类”的认知链条，正在被“AI 生成内容→网络爬虫→训练数据→新 AI 训练数据”的闭环取代。在从媒介到媒介的传递链条中，每个环节都会因模型架构差异引入噪声。因此从必然逻辑来看，大模型的“幻觉”不是程序错误，而是媒介物质性发展的代价，当我们将认知外包给超越生物限制的媒介系统时，就注定要承受其认知方式与人类经验世界方式的根本性断裂。

幻觉的必然性的另一面在于其主语并不局限于大模型，人类在认知过程中同样存在幻觉。大模型的结构本身也是参考人脑的思维方式，选择性注意重点并忽视次要信息，如深度学习模型 Transformer 通过计算词的概率分布的循环来生成文本，当其无法预测时就会自动生成“捏造”部分信息。同样，由于人本身存在既定的认知偏差，如损失厌恶、错误共识效应、刻板印象、锚定效应等，大模型也可能存在这类问题。当人们通过集体记忆来塑造历史认知时，大模型也可能正在通过训练数据来固化现有的社会偏见。

2. 数据缺陷、算法缺失与场景错位

从数据层面来看，训练数据作为大模型的信息输入，构成其媒介物质性的基础。任何媒介都依赖于其物理载体，如书籍依赖纸张，广播依赖电波，而大模型依赖的是数据集合与存储介质。大模型

训练的语料库来自人类的数字表达,其中包含系统性偏见,例如 GPT-3 数据集并未公布训练数据集的大小和内容。在关于 GPT-3 的性别偏见的调查中,立法者、银行家或名誉教授等受教育程度较高的职业以及需要艰苦体力劳动的职业,如泥瓦匠、工程师和警察都严重倾向于男性。女性更可能从事的职业包括助产士、护士、接待员、管家等。并且在性别共现测试中,最具偏见的男性的共现词为"大多数(mostly)、极好的(fantastic)、保护(protect)、生存(survive)",而女性的共现词则为"淘气(naughty)、随和(easy-going)、怀孕(pregnant)、华丽(gorgeous)"。[20]数据筛选机制的不完善,让大模型在内容生成时会再现与再生产数据中既有的各种偏见或歧视,非英语语言数据集占比的不足可能导致非西方知识表征的结构性缺失。自然语言处理(Natural Language Processing,简称"NLP")预处理中的词干提取、停用词过滤等技术操作,实质是媒介物质性对语言的暴力简化。这种"清洁化"过程抹除了语言的社会情境锚点。

算法架构也具有媒介特异性,Transformer 的自注意力机制(Self-Attention)构建了独特的认知结构。图灵奖得奖者珀尔·朱迪亚(Pearl Judea)把因果推断分成三个层面:第一层是"关联";第二层是"干预";第三层是"反事实推理"。目前的机器学习只处于第一层,只是能够关联的"弱人工智能",要实现"强人工智能"还需要具备干预和反事实推理等能力。[21]11-20 而干预和反事实推理意味着大模型需要更深的理解力和想象力去理解因果关系,涉及更高层次的因果推断。在 Transformer 的核心技术自注意力机制的运作中,通过计算每个词与其他词的关联权重来提取特征,本质上与动态的语义网络类似,如"华为"与"手机"权重高。然而自注意力机制也可能存在关联陷阱和过度拟合的问题,如模型通过共现频率学习"闪电→雷声"的强关联,但无法理解其间存在速度差异的物理因果。这种基于统计的关联捕获,导致生成文本时可能出现"先闻雷声后见闪电"的反物理叙述。因此,算法关系的多样性与精深度的缺失容易出现由于算法缺失所造成的幻觉。

大模型幻觉的另一个原因是交互界面的再中介化,主要表现在训练场景与应用场景的错位。在时空锚点上,人类通过具身化的情境认知来决策,而大模型却是去语境化的统计建模。一方面,大模型训练数据往往是从特定领域、任务或环境中收集而来,在特定的数据集上训练可以习得特定的模式与规律。另一方面,不同的应用场景涉及不同的任务要求和目标。[22]当应用场景与训练模型所基于的场景存在较大差异时,训练模型时所优化的目标函数可能并不适用于新的应用场景,从而导致模型可能无法准确地适应新情况,在处理新数据时出现偏差。此外,当用户提供的输入信息不完整或模糊时,即"提示模糊"时模型可能无法准确地将训练场景与应用场景相匹配,从而产生幻觉。

三、隐患与互补:大模型幻觉影响的多重面向

大模型幻觉并非单纯的技术缺陷,而是在数字媒介在物质性作用下的结构性产物。正如广播剧《外星人入侵地球》带来的真实恐慌,大模型生成的内容以一种更为隐性的方式侵入人的生活世界,如果机器的逻辑训练无法在短时间内提升,验证、交叉验证和第三方核查将成为日常的媒介实践,否则大模型幻觉可能会造成个体认知萎缩、社会认知分化,并在人机互动中损坏人机信任,进而破坏人机合作加剧人机冲突。

(一)负面效应:认知萎缩、社会分化与人机信任危机

1. 认知萎缩:大模型幻觉的语料混乱、结构缺陷造成认知萎缩

认知萎缩(cognitive atrophy)原是指大脑体积缩小导致的认知功能障碍,后有学者提出由于过度依赖 AI 引起的认知萎缩意味着核心认知技能的下降,尤其是个体过度依赖 AI 获取信息与处理信息会导致批判性思维、分析敏锐度和创造力的下降。[23]大模型作为一个自动内容生产机器,以其个性化与交互性容易让个体在信息摄入方面对其产生认知依赖。而无法消弭的大模型幻觉,可能进一步从信息处理上造成个体的认知萎缩。

在信息摄入层面,目前的智能技术的整合水平未必能达到足够高的水准,但是个性化和互动化

快速生成结果容易让人产生认知依赖。有研究者将同主题下 AIGC 与 UGC 内容进行比较发现，AIGC 发散性弱，主题集中度高，同质化严重且回答平庸笼络。[24]有别于传统信息源，大语言模型由于其交互性和个人化特性可以在短时间根据用户提问快速生成结果，容易让部分群体的知识获取变成“填食”行为，也就是获得表面答案而非深层次的理解。[25]

在信息处理方面，大模型数据要素存在固有缺陷，以及算法关系的错位或缺失可能出现不合理的逻辑关系或过度的因果推断。基于概率和模拟的自动化生产的软件系统，优先取悦人类反馈而非事实逻辑，使人类容易陷入这种系统设定的互动圈套，“事实让位于互动”预示着内容将变得廉价，真实成为一种稀缺资源。长期在混乱的语料、缺陷的结构中处理信息，个体对信息的批判性思维、分析敏锐度会不断下降。

2. 社会分化：扩大用户认知与行为的差异固化数字不平等

在传统数字不平等的分化基础上，AI 技术带来的数字不平等，不仅停留在浅层的、外在的对这些数智技术的接入分化和使用技能分化上，更表现为由这些差异带来的人类认知能力的分化上。其中，那些能够熟练接入和使用技术的强势用户、虽能接入但使用不熟练的中间用户、无法或拒绝接入技术的弱势用户之间存在明显差异。而大模型幻觉的存在，会进一步从信息获取、技术使用与社会影响等方面加剧不同群体之间的分化。

在信息获取方面，数字素养不高的弱势用户容易获取大模型由于幻觉而生成的不准确或不可靠信息。然而那些强势用户却能够运用设计智慧搭建基于问题的思维架构，根据大语言模型的输出结果进行批判反思与科学决策，更为高效地获取信息。

在技术应用方面，随着大模型的不断发展，其回答质量不仅取决于底层算法和训练数据，还取决于其接收的提示（问题表达）的有效性。由于提示词直接影响大模型表现，因此那些掌握熟练方法的强势用户能够在合理的提示下激发大模型能力，减少幻觉带来的影响。然而对大模型技术操作不熟练的弱势用户可能由于指令模糊或场景错位而受到幻觉的误导。

长此以往，大模型幻觉的存在使不同群体之间在信息获取与技术使用方面差距越来越大。在信息传播领域，能够准确识别和纠正大模型幻觉的人可能会掌握更多的话语权和影响力，而那些容易受到大模型幻觉影响的人则可能被误导，形成错误的观念和认知，进一步加剧社会分化。因此，相比于弱势用户，强势用户会在信息获取方面更为高效，并在技术使用中减少由于幻觉带来的负面影响。长此以往，强势用户会掌握更多的话语权与影响力，这将从多方面加剧社会分化。

3. 信任危机：信息失真与归责模糊会对人机信任产生冲击

不明来历的大模型幻觉可能对人机信任产生冲击。茹克尔（Zucker）从社会学视角划分出信任产生的三种重要模式：来源于过程的信任、来源于特征的信任、来源于制度的信任。[26]第一种模式的核心在于互惠，如交换经验或礼物互惠等；第二种模式侧重点在于在相似性中达成共识，如相似的经验、情感等；第三种模式指的是信任与正式的社会结构紧密相连，如专业背书、法律保障等。在人与大模型的关系中，人机信任即在已知不确定和脆弱的情况下，认为代理（agent）能帮助个体实现目标，信任水平会影响人与大模型的互动。[27]

而大模型幻觉可能从互惠价值与制度背书两个方面对人机信任产生冲击。在互惠层面，大模型幻觉提供的不准确与不可靠信息会降低个体对大模型互惠价值的感知。如果大模型在对话中无法完全符合用户指令，或事实性幻觉破坏大模型的准确性时，负面的反馈会降低个体对大模型互惠价值的感知，进而对大模型的能力失去信心。

在制度建立层面，平台声明与法律法规的模糊，加之技术的复杂性与黑箱性，使用户在面对大模型幻觉带来的损失时，难以明确责任主体，也会进一步对人机信任造成威胁。当大模型幻觉引发的错误和风险给用户带来损失时，源于制度信任的内容如平台声明、法律法规等仍然在发展过程中，难以明确责任主体是开发者、使用者还是技术本身，这种责任界定的模糊性使得人们在面对大模型幻

觉带来的损失时感到无助,也会进一步破坏人机信任。[28]

(二)正面效应:优化决策、激发创新与拓展知识边界

大模型幻觉作为一种复杂现象,不能简化为错误。与人类通过五官了解世界的方式不同,大模型通过自然语言的语义关系了解世界,它能够启动系统式思维对人类决策过程形成补充,其非传统、非预期的输出能够激发创新与创造,可以不断试探与拓展人类知识的边界。因此,大模型幻觉同时存在对人与社会产生正面效应的潜在价值。

1. 开启系统式认知以优化决策过程

信息处理的启发式-系统式模型(Heuristic-Systematic Model of Information Processing,简称"HSM")发现人类认知存在两种方式,即启发式认知和系统式认知。[29]当两个系统作用一致时,决策结果往往既遵从直觉又合乎理性,然而多数情景之下,两个系统存在竞争关系,占用较少心理资源的启发式容易获胜,这也是很多非理性偏差的根源。[30]49-81 如果大模型不存在幻觉,也就是说,如果大模型快速整合与生成内容的能力总是准确且可靠的,那么个体总是能够通过大模型提供的二手信息进行启发式认知,这样即使启发式与系统式认知会产生竞争,获胜概率也会大大增加,从而加剧个体的认知懒惰(cognitive laziness)。然而大模型幻觉意味着人工智能系统并非完美无缺,需要个体对大模型的输出内容进行批判性思考和评估,在人机交互中平衡启发式与系统式认知之间的竞争,优化决策过程。

当前,"脑腐"(Brain Rot)正在成为新时代的流行病,这一词的出现预示着当人过分依赖低质量且无价值和内涵的网络内容后,出现的精神和智力的衰退。"脑腐"也是人类逐渐被剥夺认知和思考的过程。大模型给人类带来的幻觉同"脑腐"的共同点在于其都体现了人类认知水平的衰退,批判思考和主动思考能力的下降。因此,大模型个性化、互动化、快速化地生成内容,容易让用户长期依靠二手信息进行启发式认知,陷入认知懒惰。不过,大模型幻觉的存在表明人工智能系统并非完美无缺,大模型存在的幻觉可能是技术的阿喀琉斯之踵,但却能够凸显人类推理判断和批判思维的能力。人类需要利用自己的知识、经验和判断力来纠正大模型的错误,同时也可以从大模型的输出中获取灵感和启示。这种人机合作的模式既可以促进技术的进步,也能够提高个体的决策能力。并且,与大模型幻觉共处需要不断进行批判性的思考和评估。当个体遇到大模型生成的不准确或不合理的输出时,会促使他们去思考为什么会出现这种情况,以及如何改进模型以提高其准确性。这种批判性思维的过程有助于人们提高自己的分析和判断能力,培养系统性认知。

2. 非传统非预期输出激发创新创造

熵(entropy)最早在热力学第二定律中被提出,指系统中分子的无序程度。1948 年香农将熵引入传播学领域,提出信息熵。信息熵是指一个信息系统的不确定性或无序程度。[31]在大模型的语境下,熵可被理解为模型输出的不确定性。从信息论的视角下看,大模型幻觉可以被视为一种信息熵的增加。契克森米哈赖在创新动力学探讨中引入了熵的范畴,并作了创造性发挥,认为创新还有比一般人所承认的那些驱力更根本、更重要、更强大的动力,那就是熵。[32]

大模型幻觉通常表现为具有不确定性、新颖性和意外性,这些都是信息熵增加的表现。低熵状态往往代表着稳定和有序,但这种有序会限制创新,而熵的增加可以打破传统的秩序和思维模式。高信息熵状态意味着多样性与可能性增加,即使大模型幻觉的输出是错误的、模糊的或不完整的,但是却能够促进不同元素、思想的融合与碰撞,带来更多可能性。因此,作为一种非预期的输出,大模型幻觉有望促进不同元素、思想的融合与碰撞,增加多样性与可能性,同时打破传统秩序与思维模式促进创新发展。

系统运行中的"负反馈机制"作为一种通过识别、传递和响应"目标偏差"或"不良结果"能够引导系统进行动态调节。大模型幻觉作为一种非传统、非预期的输出作为一种负反馈,能够在人机之间实现动态的自我纠错与持续进化,正如德勒兹所说:"人类所有的进步,不是从一个现成的答案到

下一个现成的答案，相反，人类的思考是在负反馈的问题场域中，试图打开更多的别样的思考向度。”[33]

3. 试探与拓展人类隐性知识的边界

作为知识生产网络的新节点，大模型内容生产可看作拓展知识来源的第三种途径，而大模型幻觉在一定程度上促使人们去深入挖掘潜在的、未被明确表达的隐性知识，并不断试探知识的边界，推动既有边界的拓展。知识是指建立在数据与信息基础之上，经过理论分析与实践验证的具有真实性的信念，广义的知识则不局限于狭义知识的客观性特质，还包括主观认知，凡是“行动者所持有的信念，无论是默契于心的，还是在话语层面上言明的，都应该看作知识”[34]316。在知识生产中，所有人类或非人类的参与者，包括非人的实体、技术系统、观念等在内，都被视为具有能动性的“行动者”，他（它）们通过相互关联的网络推动知识生产，人类和人工智能一样，都是知识生产网络中的节点。[35]

一方面，大模型幻觉在一定程度上为隐性知识的发展提供了新的契机。传统的知识来源主要是理性和信仰两种途径，理性强调通过逻辑推理、观察、实证研究等方式来获取知识。信仰涉及对超自然力量、宗教教义、价值观等的信任和接受。然而人工智能的知识生产超出人类理性和信仰能力的范畴，所以不是理性的过程也不是信仰的过程。也被称为隐性知识，即人工智能在人类理性思维能力以外生成的知识对人类来说是不可知的。隐性知识通常是难以明确表达和传递的个人经验、直觉、洞察力等知识形式，广义上可以理解为一种人类尚未通达的知识。当大模型产生幻觉时，可能会引发人们对已有知识体系的反思。这种反思促使人们深入挖掘潜在的、未被明确表达的隐性知识，以验证大模型输出内容的准确性。在这个过程中，人们可能会更加关注自身的经验、直觉和洞察力，从而将隐性知识显性化。另一方面，大模型幻觉的存在也意味着知识的构建是一个动态的过程。知识并非一成不变，而是在不断地被挑战、修正和扩展。大模型幻觉可以被视为对隐性知识的试探，促使人类不断探索未知领域，拓展知识的边界。

四、何以共生：重审大模型幻觉的边界与价值

埃卢尔在《技术社会》（*The Technological Society*）中提出，技术不仅仅是一个工具和机器的集合，它还是一个复杂的社会和文化系统，塑造了我们的价值观、信仰和行为，它有一种扩张和自主的趋势。技术对人类的挑战在于如何利用技术潜在的好处，同时尽量减少其对个人和社会的消极影响。[36]138 下文基于“人-机”协同视角，提出对幻觉率的容忍与调整应随应用场景而流动，大模型的温度参数可自定义以在信息准确性与创意性之间进行平衡与取舍。

（一）基于场景的温度参数：在准确性与创意性之间的取舍

人类的对话场域是流动而微妙的，而大模型幻觉产生的原因之一就在于其取消了对话过程中的“意义协商”，使得生成信息与场景错配。当医生无法理解 AI 诊断建议的生成逻辑，记者难以追溯自动写作系统的选题偏好，就会形成技术权威与人类经验的认知断层。因此，场景决定了对算法的特定需求，对大模型幻觉率的控制应结合具体场景进行讨论。

在语言模型中，温度参数（Temperature Parameter）是一种调整生成文本特性的指标，主要影响大模型生成文本的随机性和多样性。[37]信息熵作为衡量概率分布不确定性的一个指标，温度参数的调整实际上就是改变信息熵。当温度参数增大时，信息熵增大，生成文本表现出更高的随机性和可能的创意性；当温度参数减小时，信息熵减小，文本倾向于遵循常见的模式，准确性在一定程度上更容易保证，但可能缺乏创意。目前技术支持对大模型温度参数的自定义。因此，根据具体的应用场景和任务需求，温度参数可以随之调整。以新闻传播领域为例，在需要准确和可靠结果的场景下，温度参数可以调低以确保信息的准确性与可靠性。在侧重创意性与多样性的场景下，可以调高温度参数以增加任务的创意性与独特性。

在实际应用场景中，用户需要依赖大模型输出准确信息以进行可靠决策时，内容的准确性与可

靠性往往是首要考虑因素,尤其是在医疗、法律等领域,大模型的幻觉率应尽可能降低。在新闻生成场景中,大模型可能混合真实事件与统计幻象,编辑因无法理解其“思考”路径而丧失事实核查能力。在发布场景下,新闻媒体如果发布不准确或不可靠的信息,可能会引发公众对新闻媒体的信任危机。因此,在新闻内容报道或发布这类强调准确性与可靠性的高风险场景下,可以调低模型在相关应用场景下的温度参数,将大模型幻觉率降低以确保内容的真实与准确,减少大模型幻觉可能引致的问题与风险。

对幻觉率的容忍与调整应随应用场景而流动,用户可以自定义温度参数在信息准确性与创意性之间进行平衡与取舍。创意性是解决复杂问题的重要突破口。在面对一些没有明确解决方案的问题时,对于大模型幻觉率的容忍度可适当增大。例如在新闻选题阶段,大模型幻觉可能带来一些新颖的、意想不到的话题方向或角度。对于一些较为常规的新闻事件,大模型可能会基于其对大量信息的整合和理解,提供一些独特的解读视角或相关的衍生话题,帮助新闻工作者打开思路,挖掘出更具吸引力的新闻选题。在新闻策划方面,大模型可根据对不同主题和事件的关联分析,提供一些创新的跨领域、跨学科的想法和框架,使新闻专题更具深度和广度。因此,在新闻选题、策划这类强调创意性与独特性的低风险场景下,可以合理调高温度参数,最大程度激发大模型幻觉的潜在创新、创造价值。因此,场景决定对算法的特定需求,对幻觉率的容忍与调整应随应用场景而流动,以便在信息准确性与创意性之间进行平衡与取舍。

(二)与大模型幻觉共生:人机协同下人人都是提示词工程师

从现实应用情况而言,大模型幻觉已经彻底融入文本生产和传播的各个环节之中。在生成式 AI 应用和实践所能涵盖的范围之内,与幻觉共生的主体不再是中介机构,而是被下放到每个与之对话的用户个体。[38]提示工程(Prompt Engineering)是依托于大模型的智能能力,通过设计模型应执行任务的自然语言文本,有组织地将提问想法转化为具有一组输入集或指令集的过程。提示素养(Prompt Literacy)则是用户在这个过程中能否有效输入指令、引导 AI 优化输出最终实现高效协同的能力,包括对提示设计原则和方法的掌握、具备评估和优化提示的能力以及批判思维和创新意识的养成等。[39]随着大模型越来越复杂,其回答质量不仅取决于底层算法和训练数据,还取决于其接收的提示(问题表达)的有效性,未来人人都将直接或间接地成为提示语工程师。[40]

大模型幻觉作为一种复杂现象,除了技术本身,同时也与用户“提示素养”相关。一方面,相比于提升素养不高的用户,素养高的用户更能够根据场景合理设计提示词,避免由于指令模糊、场景错误而产生忠实性幻觉。或凭借自己的知识对大模型的错误输出进行纠正,通过质疑和逻辑分析,筛选出大模型输出中的不合理内容,避免被误导。另一方面,提示素养高的群体能更理解大模型的局限性与适用范围,根据应用场景更灵活地自定义不同温度参数,在使用过程中挖掘出大模型幻觉的潜在价值。

因此,降低由于指令模糊、场景错误而产生幻觉负面效应的可能性,用户应提高“提示”素养,并在人机协同与合作中挖掘大模型幻觉的潜在价值。如何提升“提示素养”?在学习方面,用户应深入理解模型的训练方法、数据来源、知识范围及风险局限等,以便在设计提示时更好地把握模型的性能和特点。在实践方面,可以在积极与大模型交互中尝试不同类型的任务和问题,在实践中不断提升设计提示的有效性。此外,对输出结果应保持批判性思维,学会分析评估与追问验证。因此,未来人人都是提示词工程师,指令蕴含的知识越丰富,组织越完善,模型求解复杂任务的能力也会越强,人机协作依然是应对大模型幻觉问题的最佳方式。一方面,用户应提高“提示”素养,理解大模型致幻的边界,并接受幻觉率在不同场景的适配程度,降低由于指令模糊、场景错误而产生幻觉负面效应的可能性;另一方面,不断优化、创新指令有望在人机协同与合作中不断发掘大模型幻觉的潜在价值。

五、结语:在人机协同中实现大模型幻觉价值最大化

技术的价值呈现本身具有很强的情境依赖性,而人作为技术应用的主体,决定了这种情境的塑

造以及技术价值在不同社会情境下的具体展现。[41]随着大模型不断嵌入各行各业与日常生活,大模型幻觉已经成为影响技术发展、人机信任的重要问题。然而不透明性、概率性与自主性作为人工智能系统的本质特征,难以使大模型幻觉完全消弭,但讨论大模型幻觉问题不能只关注其消极面向,更应该在促进技术发展与人机协同的语境下,重新审视 AI 大模型幻觉的边界与价值。人与媒介和世界关系的再中介和断裂在深层次上呼应了海德格尔对技术“座架”(Gestell)本质的批判,技术媒介正在重塑人类的存在方式,而我们对这种重塑的后果尚未做好认知准备。

笔者从大模型技术底层逻辑出发,分析了大模型幻觉的表现特征与产生机制,认为作为复杂系统的大模型由数据要素与算法关系组成,幻觉产生的根本原因在于算法无法在适配场景下将要素进行合理的关系连接;本研究同时认为,需辩证看待大模型幻觉可能带来的消极与积极效应。一方面,大模型幻觉中语料混乱、结构缺陷可能会对用户认知造成干扰,用户对幻觉的认知与行为差异会进一步加大数字不平等,并且幻觉容易造成人机信任危机;另一方面,与人类通过五官了解世界的方式不同,大模型通过自然语言的语义关系了解世界,能够启动系统式思维对人类决策过程形成补充,其非传统、非预期的输出能够激发创新与创造,不断试探与拓展人类知识的边界。

需要注意的是,“大模型幻觉”一词可能还存在伦理上的污名化问题,这一比喻将人工智能的负面问题与精神疾病相关联,因此亟须正视大模型幻觉的诸多影响。大模型幻觉作为一种难以消解的复杂现象,对我们的挑战在于如何利用其潜在好处,同时尽量减少其对个人与社会的消极影响。“接受大模型幻觉永远无法被消弭的事实,意味着我们需要重新考虑何时、何地以及如何使用大模型。它们是很棒的创意创造者,但无法独立解决问题。因此可以把它们放到一个有验证者的架构中来利用它们。”[11]场景决定对算法的特定需求,对幻觉率的容忍与调整应随应用场景而流动,用户可以自定义温度参数在信息准确性与创意性之间进行平衡与取舍。未来,人人都是提示词工程师,用户应提高“提示”素养,降低幻觉可能带来的负面效应,并在人机协同中不断发掘大模型幻觉新的价值。

参考文献:

[1] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models. (2021-10-06) [2024-09-10]. http://arxiv.org/abs/2108.07258.

[2] 闫坤如. 人工智能价值对齐的价值表征及伦理路径. 伦理学研究, 2024(4): 94-100.

[3] 张静, 陈方迪. 2023 年度词都和 AI 相关: 剑桥词典选“幻觉”, 韦氏词典选“真实”. 澎湃新闻. (2023-11-29) [2024-09-29]. https://www.thepaper.cn/newsDetail_forward_25455617.

[4] HUANG L, YU W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. (2023-11-19) [2024-09-29]. https://doi.org/10.1145/3703155.

[5] KALAI A T, VEMPAL S S. Calibrated language models must hallucinate. (2024-03-20) [2024-10-29]. https://doi.org/10.48550/arXiv.2311.14648.

[6] DAHL M, MAGESH V, SUZGUN M, et al. Large legal fictions: profiling legal hallucinations in large language models. Journal of Legal Analysis, 2024, 16(1): 64-93. [2024-06-21]. https://doi.org/10.1093/jla/laae003.

[7] AGRAWAL G, KUMARAGE T, ALGHAMDI Z, et al. Can knowledge graphs reduce hallucinations in LLMs: A Survey. (2024-03-16) [2024-10-29]. http://arxiv.org/abs/2311.07914

[8] 张铮, 刘晨旭. 大模型幻觉: 人机传播中的认知风险与共治可能. 苏州大学学报(哲学社会科学版), 2024(5): 171-180.

[9] 彭兰. 从 ChatGPT 透视智能传播与人机关系的全景及前景. 新闻大学, 2023(4): 1-16+119.

[10] HOHWY J. The predictive mind. Oxford: Oxford university press, 2013.

[11] KAMBHAMPATI S. Can large language models reason and plan? Annals of the New York academy of sciences, 2024, 1534(1): 15-18. [2024-03-26]. https://doi.org/10.1111/nyas.15125.

[12] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation. ACM computing surveys, 2023, 55(12): 1-13. [2024-07-14]. https://doi.org/10.1145/3571730.

[13] RAWTE V,CHAKRABORTY S,PATHAK A,et al. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Proceedings of the 2023 conference on empirical methods in natural language processing,2023. [2023-10-04]. https://doi. org/10. 48550/arXiv. 2310. 04988.
[14] 梅夏英. 复杂系统与智能涌现:未来数字法研究的范式图景. 法学家,2024(5):45-59+191-192.
[15] 江积海,阮文强. 新零售企业商业模式场景化创新能创造价值倍增吗? 科学学研究,2020,38(2):346-356.
[16] 严昊,刘禹良,金连文,等. 类 ChatGPT 大模型发展、应用和前景. 中国图象图形学报,2023,28(9):2749-2762.
[17] POHL R F. Cognitive illusions:Intriguing phenomena in thinking,judgment,and memory. Oxfordshire:Taylor & Francis Group,2016.
[18] 杜文静. 法律人工智能概率推理的困境与破解. 学术研究,2022,45(4):29-34.
[19] 方婕妤. 人类中心主义与媒介物质性之间的抉择——论唯物现象学的内在矛盾. 新闻界,2025(1):87-96.
[20] BROWN T,MANN B,RYDER N,et al. Language models are few-shot learners. Advances in neural information processing systems,2020,33(4):1877-1901. [2020-07-22]. https://doi. org/10. 48550/arxiv. 2005. 14165.
[21] 朱迪亚,王白尔,麦肯齐. 为什么:关于因果关系的新科学. 江生,于华,译. 北京:中信出版集团股份有限公司,2019.
[22] 许志伟,李海龙,李博,等. AIGC 大模型测评综述:使能技术、安全隐患和应对. 计算机科学与探索,2024,18(9):2293-2325.
[23] MUSHESHE M,CHRISPO M. A conceptual study of cognitive atrophy in homo sapiens through a darwinian analysis of overreliance on artificial intelligence. East African journal of interdisciplinary studies,2025,8(2):185-207. https://doi. org/10. 37284/eajis. 8. 2. 3463.
[24] 王浩伟,汪璠,王秉琰. 主题视角下生成式人工智能生成内容与用户生成内容的比较. 情报理论与实践,2023,46(10):200-207+199.
[25] 彭兰. 智能生成内容如何影响人的认知与创造? 编辑之友,2023(11):21-28.
[26] ZUCKER L G. Production of trust:institutional sources of economic structure. Research in organizational behavior,1986(8):53-111.
[27] 齐玥,陈俊廷,秦邵天,等. 通用人工智能时代的人与 AI 信任. 心理科学进展,2024,32(12):1-13.
[28] 张洪忠,任吴炯. 超越"第二自我"的人机对话——基于 AI 大模型应用的信任关系探讨. 新闻大学,2024(3):47-60.
[29] 孙彦,李纾,殷晓莉. 决策与推理的双系统——启发式系统和分析系统. 心理科学进展,2007(5):721-726.
[30] GILOVICH T,GRIFFIN D W,KAHNEMAN D. Heuristics and biases:the psychology of intuitive judgement. Cambridge:Cambridge university press,2002.
[31] 林爱珺,陈亦新. 信息熵、媒体算法与价值引领. 湖南师范大学社会科学学报,2022,51(2):125-131.
[32] HARDY D. Creativity:Flow and the psychology of discovery and invention. Personnel Psychology,1998,51(3):794-797.
[33] 姜宇辉. "相信此世"——在与克尔凯郭尔的对话之中重释德勒兹的"否定主义". 四川大学学报(哲学社会科学版),2023(5):135-145+197.
[34] 吉登斯. 社会的构成:结构化理论大纲. 李康,李猛,译. 北京:中国人民大学出版社,2016.
[35] 刘大年,曹月. 知识的幻象:人工智能与知识变迁. 现代出版,2024(9):53-67.
[36] ELLUL J. The technological society. New York:Alfred A. Knopf,1964.
[37] KALAI A T,VEMPALA S S. Calibrated language models must hallucinate. (2024-03-20)[2024-10-29]. http://arxiv.org/abs/2311.
[38] 经羽伦,张殿元. 生成式 AI 幻象的制造逻辑及其超真实建构的文化后果. 山东师范大学学报(社会科学版),2024,69(5):113-126.
[39] 赵晓伟,祝智庭,沈书生. 教育提示语工程:构建数智时代的认识论新话语. 中国远程教育,2023,43(11):22-31.
[40] CLARK P A. AI's rise generates new job title:prompt engineer. (2023-02-22)[2024-09-29]. https://publicservicesalliance. org/wp-content/uploads/2023/02/AIs-rise-generates-new-job-title-Prompt-engineer. pdf.
[41] 叶继红,雷德森. 科学技术的社会变迁:一个社会学的分析. 科学管理研究,2004(5):51-55.

Manifestation Characteristics, Controversies over Effects, and Potential Values of Large Language Model Hallucinations

Yu Guoming, *Jin Liping* (Beijing Normal University)
Su Fang (University of International Relations)

Abstract: Hallucinations within large language models have become deeply entrenched in every phase of content production and dissemination processes, thereby emerging as a crucial problem that exerts a significant impact on individual cognitive processes, social stratification, and human-machine trust relationships. Nevertheless, considering that opacity, probabilistic nature, and autonomy constitute the intrinsic characteristics of artificial intelligence systems, it is a difficult task to entirely eradicate hallucinations in large language models from a technical standpoint. This paper commences from the fundamental logic underpinning large language model technologies. It conducts an in-depth analysis of the definitional implications, expressive features, and generative mechanisms of hallucinations within large language models. Moreover, it adopts a dialectical approach to examine both the negative ramifications and potential values of these hallucinations. As a complex and intractable phenomenon, the primary challenge presented by large language model hallucinations for us lies in the optimization of leveraging their potential advantages while concurrently minimizing their adverse effects to the greatest extent possible. Consequently, the tolerance levels and adjustment strategies regarding the hallucination rate should be adaptable to diverse scenarios. Users are enabled to customize temperature parameters, thereby facilitating the achievement of a balance and enabling trade-offs between information accuracy and creativity. Additionally, users are required to enhance their proficiency in "prompting" to mitigate the potential negative consequences induced by hallucinations and to maximize the potential value of large language model hallucinations within the context of human-machine collaboration.

Key words: large language model; hallucination of intelligence; human-computer collaboration; value scenarios; temperature parameter

---

■**收稿日期**:2025-05-07
■**作者单位**:喻国明,北京师范大学新闻传播学院;北京 100875
金丽萍,北京师范大学新闻传播学院传播创新与未来媒体实验平台
苏 芳,国际关系学院国家安全学院;北京 100091
■**责任编辑**:刘金波