生成式人工智能虚假信息传播风险评价指标体系构建

魏畅畅，李 婵

（福建农林大学公共管理与法学院，福州 350002）

摘要： 生成式人工智能技术在显著提升信息生产效率的同时，也因其技术特性加剧了虚假信息的生成与传播，引发社会不稳定问题。以风险管理理论为基础，系统梳理生成式人工智能背景下虚假信息传播的治理现状，剖析其传播特点与社会影响；结合信息传播理论核心框架，从传播主体、传播内容、传播媒介、传播受众、传播效果5个维度，构建生成式人工智能虚假信息传播风险评价指标体系，并运用层次分析法确定各层级指标权重。研究最终构建出包含27项具体指标的完整评价体系。结果显示，传播内容风险在指标体系中权重居首，其次是传播效果风险、传播媒介风险等。研究成果可为精准识别、科学评估生成式人工智能虚假信息传播风险提供有益参考和指导。

关键词： 虚假信息传播风险；生成式人工智能；评价指标体系

中图分类号： G202；TP309.2 **文献标志码：** A **文章编号：** 1671－1807(2026)09－0263－09

近年来，人工智能（artificial intelligence，AI）领域有了突破性的进展，ChatGPT“火爆出圈”，标志着人工智能发展进入新阶段，市场上涌现出了各种类型的生成式人工智能。然而，随着人工智能技术的不断发展和更迭，其潜在风险逐渐暴露于公众视野中。由于其无法真正推理，易生成看似合理的错误信息，且普通大众难以辨别，加剧了虚假信息传播风险。这一风险会在政治层面扩大立场分歧、社会层面冲击信任体系，模糊真相与谎言边界。因此，对生成式人工智能产生的虚假信息传播的风险进行深入分析和治理研究是十分重要的。

当前，学界认为，生成式人工智能虚假信息的传播风险是技术特性[1]、社交环境[2]等多因素交织作用的结果。在技术因素层面，Tortora[3]明确指出，机器学习与深度学习技术赋予生成式人工智能强大的多模态内容生成能力，可依据各类提示生成看似新颖且贴合人类创作逻辑的文本、图像、音频、视频等内容，其中生成式对抗网络等算法的应用，进一步降低了虚假内容的创作门槛，使其在形式上难以被直观甄别[4]。贾开和薛澜[5]则从技术本质切入，提出人工智能自身技术逻辑的复杂性与应用过程的模糊性。周妍和沈天健[6]进一步界定了此类虚假信息的核心属性，明确其是经由生成式人工智能大模型生产的内容，区别于传统人工编造的虚假信息，具有生成效率高、拟真度强、传播速度快的鲜明特征，这一属性从根源上加剧了风险的扩散潜力。在社交环境因素层面，社交平台的算法逻辑与用户行为特征为虚假信息提供了滋生土壤。研究发现，社交平台的个性化推荐算法构建了封闭的信息茧房，使虚假信息能够精准匹配目标受众的认知偏好与情绪诉求。生成式人工智能彻底重构了虚假信息的传统传播模式，虚假信息的广泛传播最终会衍生出一系列社会风险，核心表现为社会信任消解与公共秩序干扰，且在不同场景下呈现差异化危害。

针对上述风险，国内外学者围绕全生命周期治理、多主体协同等核心思路，探索形成了多元化治理策略，普遍认为治理需实现法规、技术、社会力量的协同联动。生成式人工智能虚假信息治理需贯穿数据采集、模型开发、应用部署、内容传播的全生命周期[7]，依据应用场景制定差异化方案[8]。在数据采集阶段，治理重点在于强化源头审核与隐私保护；在模型开发阶段，建立恶意查询黑名单，限制有害内容生成。同时，学界共识表明，单一主体的治理难以应对跨场景、跨地域的虚假信息风险，需构

收稿日期： 2026-01-14

基金项目： 国家社会科学基金（24CTQ040）

作者简介： 魏畅畅（1998—），女，河南商丘人，硕士研究生，研究方向为公共政策；通信作者李婵（1985—），女，山东潍坊人，博士，副教授，研究方向为信息资源管理、数字风险管理。

建政府、企业、公众协同联动的治理格局。政府需发挥监管主导作用，企业需强化主体责任，公众需提升数字素养。

综上，学界已围绕生成式人工智能虚假信息传播风险与治理形成了多维度研究框架，明确了风险的形成机制与治理的核心路径。但是，关于该方面风险量化评估体系尚不健全，缺乏科学、可操作的层级化风险指标工具，致使对不同应用场景下风险的差异化实证分析缺乏支撑。基于上述思考，现以生成式人工智能虚假信息为研究对象、以风险治理为视角研究生成式人工智能虚假信息传播风险问题，结合信息传播理论构建生成式人工智能虚假信息传播风险评价指标体系，运用层次分析法构建层级模型，科学确定各级指标权重，并依据指标的重要性程度对生成式人工智能虚假信息传播风险管理提供对策建议，对虚假信息传播风险的防范具有重要的理论和实践价值。

1 生成式人工智能虚假信息传播风险概述

1.1 生成式人工智能虚假信息传播风险的特点

生成式人工智能虚假信息传播具有极强的隐匿性[9]。作为人工智能的分支领域，它依托机器学习与深度学习技术，可生成文本、图像、音频、视频等类人化内容，并通过自回归生成模型实现与用户的自然流畅交互。这种隐蔽的传播模式消解了用户的戒备心理，掩盖了观点操控者的真实意图；部分AI应用更采用虚实交织的话术，在回复中掺杂真实信息，使用户产生“获取有效知识”的错觉。随着用户对AI的依赖加深，错误观点会被持续渗透，最终潜移默化地影响用户的认知与判断。

生成式人工智能虚假信息传播风险具有复杂性高的特点[10]。生成式人工智能传播风险的复杂性，集中表现为其彻底颠覆了传统独立、单层、定向的风险传导模式。依托机内交互、人机交互与人人交互构成的复杂网络结构，风险路径呈现出联动交织、层次多元、方向多维的显著特征。这种风险传导机制不再是简单的线性延伸，而是形成了相互关联、层层嵌套的立体网络，使得风险源头隐匿于错综复杂的交互链路之中，风险传导路径在多维度交互中愈发模糊难辨，风险影响的范围与程度也因交互网络的动态演变而难以精准测度。

生成式人工智能虚假信息传播风险具有强渗透力的特点。生成式人工智能广泛应用于社会各领域[2]，虚假信息借助智能交互、内容生成等功能，以看似合理的形态融入日常信息传播链条。无论是用户获取知识、进行决策，还是参与社交互动，都可能在不经意间受到虚假信息的影响，其传播风险已悄然渗透到社会运行的每一个环节，给信息安全、社会秩序带来潜在威胁。

1.2 生成式人工智能虚假信息传播的影响

生成式人工智能虚假信息的肆意传播在社会、政府、个人等方面产生了负面影响。生成式人工智能的本质是信息，其影响社会的底层基础依然是传播问题，当生成式人工智能深嵌于智能传播格局之中，互联网信息秩序在技术促动下发生巨大转向，对信息传播全过程产生颠覆性影响。

在社会层面，生成式人工智能虚假信息的传播可能引发公众恐慌，对社会信任体系造成严重破坏。虚假信息往往具备显著的深度伪造特征[11]，这种特性使其极易对社会秩序产生深层次、广泛性的负面影响。以及虚假信息往往会迎合特定群体的观点和利益，可能会导致不同群体之间的认知差异进一步扩大，引发群体之间的对立和冲突，加剧社会分化和群体冲突。在政府层面，虚假政治新闻的影响比关于恐怖主义、自然灾害或金融信息的虚假新闻更明显，政治假新闻对当代社会构成持续威胁，对公共和民主机构产生负面影响，虚假信息可被用于操纵舆论、影响选举等政治活动。在个人层面，生成式人工智能采用基于人类反馈的强化学习，在很多时候可以以假乱真，让使用者觉得是跟真人对话，弱化了人类主体地位，会产生因信息失当传播而带来的负面影响。错误的信息会损害认知，促进错误信念的形成，从而影响推理和决策。

1.3 生成式人工智能虚假信息传播风险治理现状

生成式人工智能的迅猛发展，为信息生产与内容创作带来了前所未有的便利，极大提升了内容产出的效率与便捷性。然而，用户以极低成本、极高效率批量生成具有高度逼真性与迷惑性虚假信息，这类虚假信息往往兼具强烈的误导性与煽动性，更关键的是，一旦被公众接受，其误导效应便会呈现出极强的顽固性与持续性，常规辟谣手段难以实现彻底消解，最终对公众认知造成难以逆转的负面影响[12]。在应对生成式人工智能虚假信息的过程中，各利益相关方如政府、企业、公众等之间存在着复杂的利益与目标冲突，这使得任何单一主体的治理模式都难以有效应对[13]。现有的单一治理模式未能有效调动多方资源进行协作，导致各方往往陷入各自为战的困境，降低了整体治理效能[14]。因此，

建立一个能够有效整合政府、企业、社会等多方力量,成为化解生成式人工智能虚假信息传播风险、突破当前治理困境的关键路径与必然选择。

生成式人工智能虚假信息治理过程中,政府通过监督和协调发挥主导与引领作用,通过制定宏观战略、法律法规与行业标准,为治理工作设定清晰的规则框架,并引导社会形成对技术的科学认知。政府通过出台专项法规与国家标准,明确 AI 内容标识要求;实施算法与模型备案制度,强化源头管控;同时结合法治与激励性监管,监督平台履责,推动企业落实合规要求。企业承担关键执行与内部监督责任,作为技术研发与应用的核心主体,企业需将治理要求转化为具体的技术方案与平台规则落到实处。企业通过建立全平台主动监测机制,利用智能算法实时筛查风险信息[15];引入第三方机构独立评估模型合规性,公开结果接受监督;快速启动风险处置程序,对核实虚假内容采取下架、警示等,防止风险扩散。社会履行广泛参与外部监督职能,社会力量应通过多元渠道参与政策讨论,反映公众诉求;同时,作为重要的制衡与补充力量,对技术与平台的应用进行监督,推动治理的公开透明。社会为推动生成式 AI 虚假信息多元共治,联动多方力量开展线上线下融合的科普教育,增强公众风险辨识能力;社会组织搭建物质奖励、荣誉表彰等立体化激励框架,激活社会监督力量,推动治理模式向全社会协同延伸。各主体协同治理如图 1 所示。

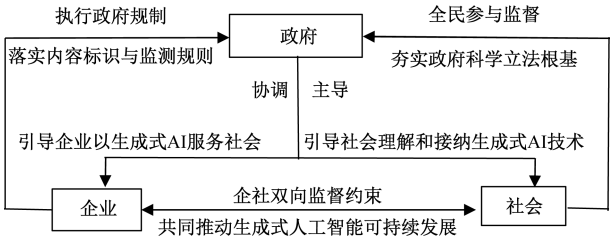


图 1　政府-企业-社会协同治理

2　生成式人工智能虚假信息传播要素风险源分析

拉斯韦尔 5W 模式是传播学研究中的基础理论框架,由美国学者哈罗德·拉斯韦尔于 1948 年提出。该模式通过回答 5 个基本问题来分析和评估传播过程的结构与特征,即传播主体、传播内容、传播渠道、传播受众以及传播效果。运用该理论对生成式人工智能虚假信息传播的要素风险来源展开拆解,精准定位风险演化的关键节点,最终确定虚假信息传播主体、受众、内容、媒介、效果为 5 大评价维度,精准定位各环节的核心风险源。

2.1　传播主体风险

生成式 AI 虚假信息的生产与传播呈现多元主体共谋特征,其传播主体风险作为源头性风险,涵盖算法偏见、技术滥用、动机隐匿及传播者教育水平与 AI 依赖度差异等多维度因素。

(1)算法偏见程度[16]。指生成式 AI 虚假信息传播平台的推荐算法因设计或训练偏差,对虚假信息产生倾向性推荐的程度,核心反映技术主体对虚假信息的助推能力。该指标可通过同期平台虚假信息平均推荐量占所有内容平均推荐量的比例衡量,算法偏见会推动虚假信息突破初始传播范围。

(2)技术工具的滥用频率[17]。指生成式 AI 虚假信息的制作、传播工具被恶意主体违规使用的次数。生成式 AI 作为生产端可批量生成高仿真虚假内容,传播端则借助各类渠道实现广泛扩散。该指标通过统计生成与传播工具的滥用次数衡量,频率越高,虚假信息的传播效率与危害程度越强。

(3)动机行为的隐匿性[18]。指恶意传播主体通过技术手段隐匿身份、轨迹与动机,利用虚拟身份发布误导信息实施诈骗。测量动机行为隐匿性需结合量化与质性方法,实现多维交叉验证。具体表现为通过隐蔽情境模拟与李克特 7 点量表量化外显行为与动机的偏离值,再以半结构化深度访谈进行校准,有效排除无意偏差,实现多维度交叉验证。

(4)受教育水平差异[19]。指不同受教育水平的传播主体在辨别生成式 AI 虚假信息能力上的差距。受教育水平较低的群体因缺乏 AI 技术认知和逻辑判断能力,易误信并转发虚假信息,进而推动虚假信息的分层传播与范围扩大。

(5)对 AI 的依赖程度[20]。传播主体对生成式 AI 的依赖程度越高,越易形成认知偏差,尤其在专业领域因知识局限过度依赖 AI,导致虚假信息信任门槛降低、传播加速、危害扩大。

2.2　传播内容风险

生成式人工智能虚假信息传播内容风险是连接传播主体与传播受众的核心载体,通过虚假内容生成的频率、AI 提示错误覆盖率、信息来源模糊度、逻辑矛盾占比、热点场景贴合度、内容更新速率作为衡量其指标层指标。

(1)虚假内容生成的频率[21]。指传播主体借助 AI 工具批量制造虚假内容的频率,频率越高,流向

受众的虚假内容越多，受众接触到虚假信息的概率越大。

(2)AI提示错误覆盖率。传播主体通过提示词设计缺陷或恶意引导，导致AI输出内容包含错误信息的比例，如模糊提示夸大某成分效果，那么AI就会输出错误内容，且错误覆盖范围随提示词复用率提升而扩大，这直接影响受众接收的内容真实性。

(3)信息来源模糊度。指虚假内容中信息来源未标注或伪造的比例，以未标注或伪造来源的虚假内容数量占比衡量。占比越高，内容可信度造假嫌疑越大，误导受众风险越高。

(4)逻辑矛盾占比。指虚假内容中存在时间线混乱、数据冲突、因果矛盾等逻辑问题的比例。矛盾占比越低，内容逻辑越自洽，受众被误导的概率越高。

(5)热点场景贴合度[22]。指虚假内容与平台热点、受众关注场景的贴合比例，是影响其扩散效率的关键。贴合度越高，虚假信息越易契合算法偏好，触达受众规模越大。

(6)内容更新速率[23]。指传播主体为规避审核，利用生成式AI快速产出多版本虚假内容的频率。该指标通过统计同一虚假主题的AI内容版本数量衡量，迭代速率越快，规避审核的概率越高。

综上所述，生成式人工智能虚假信息传播内容指标层的指标选定为虚假内容生成的频率、AI提示错误覆盖率、信息来源模糊度、逻辑矛盾占比、热点场景贴合度、内容更新速率。

2.3 传播媒介风险

生成式人工智能虚假信息传播媒介风险是指传播媒介在承载、扩散生成式人工智能虚假信息过程中，因自身特性或管控不足引发的风险。

(1)媒介覆盖规模。指平台的用户总量，是生成式AI虚假信息风险扩散的基础载体[24]。规模越大，虚假信息越易从小众传播进入大众视野，AI批量生成能力进一步放大传播优势，覆盖规模超标会导致初始传播失控。

(2)信息扩散速率。指生成式AI虚假信息在社交及短视频平台的传播效率，通过单位时间内传播节点数量衡量[25]。速率越快，虚假信息越易在官方辟谣前完成大范围渗透，后续管控成本越高。

(3)用户互动便捷性。指平台点赞、评论、转发等操作的门槛，通过用户完成互动的步骤数衡量[26]。便捷性过高会激活裂变机制，使虚假信息扩散至私域等隐蔽场景，加剧传播风险。

(4)跨媒介迁移便捷性。指生成式AI虚假信息通过多模态内容与媒介便捷操作在不同平台间的传播能力，适配性越强，越易规避单一媒介管控，实现跨平台扩散。

(5)审核人员专业能力。指平台审核人员对AI虚假内容的识别水平，可通过接受专业培训的人数衡量[27]。能力不足会导致风险拦截失效，损害公信力并增加后续管控成本。

(6)违规内容处置效率。指平台对已发现虚假信息的处理能力，以从发现到删除的耗时衡量[28]。效能越低，虚假信息越易在处置前扩散，且发布者可能通过跨平台迁移反复传播，扩大危害范围。

综上所述，生成式人工智能虚假信息传播媒介的指标层的指标选定为媒介覆盖规模与密度、信息扩散速率、用户互动便捷性、跨媒介内容迁移便捷性、审核人员专业能力、违规内容处置效率。

2.4 传播受众风险

生成式人工智能虚假信息传播受众风险是指受众在接触、认知、互动及再传播生成式人工智能虚假信息过程中，因自身认知特征、行为习惯或参与方式，导致虚假信息风险放大的现象。

(1)信息认知素养水平。指受众辨别、评估生成式AI虚假信息的核心能力[29]。素养越低，受众越易被高仿真AI虚假内容误导，不仅自身蒙受损失，还可能因主动转发进一步放大虚假信息传播风险。

(2)二次传播频率。指受众将虚假信息转发、分享给他人的频率，反映其对虚假信息的扩散助推能力。频率越高，受众从信息接收者转变为风险扩散者的概率越大，风险越强。

(3)虚假信息参与度。指受众对生成式AI虚假信息的互动行为参与强度，可通过点赞量、评论量等数据衡量。参与度越高，虚假信息越易引发受众共鸣、吸引新受众关注，还会诱导新受众产生“多人关注即真实”的认知偏差，进一步放大传播受众风险。

(4)虚假信息关注度。指受众对生成式AI虚假信息的注意力投入程度，通过浏览时长、重复浏览次数等衡量。关注度过高会导致注意力被劫持，形成长期认知偏差，增加纠正难度。

(5)受众触达深度。指生成式AI虚假信息渗透至受众家庭群、同事群等核心圈层的程度。此类虚假信息常依托熟人信任降低辨别门槛，在核心圈层内的传播效率高、误导性强，易造成风险在高信

任度场景中集中扩散。

综上所述,生成式人工智能虚假信息传播受众的指标层的指标选定为信息认知素养水平、二次传播频率、虚假信息参与度、虚假信息关注度、受众触达深度。

2.5 传播效果风险

生成式人工智能虚假信息传播效果风险是指生成式人工智能虚假信息在传播后,对受众认知、社会行为、公共秩序及政策执行等领域产生的实际危害与负面影响程度。

(1)错误认知覆盖广度。指生成式 AI 虚假信息致使受众形成错误认知的人群范围,通过直接与间接接触人群的错误认知占比衡量[30]。覆盖广度越大,对社会认知基础的冲击范围越广,后续纠正的人群基数也越大。

(2)认知偏差固化深度。指生成式 AI 虚假信息造成的错误认知在受众脑中的扎根牢固程度,认知偏差固化深度的测量以短期与长期错误认知留存率的算术平均值为核心量化依据[31]。先标定虚假信息标的、构建判定量表及标准;再双节点测算留存率并取均值;最后通过真实信息矫正测试、科学认知接受/抵制度测评补充验证,排除假性高固化,保障结果可靠。固化程度越深,受众越难接受真实信息,甚至抵制科学认知,易引发长期认知扭曲与个体健康危害。

(3)认知纠正难度。指消除生成式 AI 虚假信息所致错误认知的成本与效率,其高仿真特性会降低纠正效率、拉长纠正周期、提升纠正难度[32]。该指标可通过纠正所需人力与财力成本衡量,难度越大意味着认知后果越严重、社会修复成本越高。

(4)群体非理性行为强度。指生成式 AI 虚假信息引发的群体非理性行为的规模与激烈程度,以参与人数占区域总人口的比例衡量,比例越大风险越强[33]。其强度越高,对局部乃至区域社会秩序冲击越直接,易打乱正常生产生活节奏、破坏社会稳定。

(5)公共资源消耗成本。指政府、社会组织等为应对生成式 AI 虚假信息风险投入的人力、财力、物力总和,以 3 类资源实际投入核算总消耗成本,成本越高风险越强[34]。其越高,越会挤占医疗、教育等公共服务资源,导致供给不足、质量下降,影响民众正常生活。

(6)虚假信息舆论占比。指某公共话题总舆论中生成式 AI 虚假信息的占比,以虚假信息数量与总舆论信息数量的比值衡量,占比越高风险越强[35]。占比越高,舆论场越失真,越易误导公众形成片面认知,干扰舆论理性走向。

(7)舆论极化程度。指导致公众观点非黑即白分裂、多元观点趋极端[36],可通过极端观点与理性讨论占比衡量,极化程度越深,极端言论越多,对舆论生态、社会矛盾及凝聚力的负面影响越显著。

综上所述,生成式人工智能虚假信息传播受众的指标层的指标选定为错误认知覆盖广度、认知偏差固化深度、认知纠正难度、群体非理性行为强度、公共资源消耗成本、虚假信息舆论占比、舆论极化程度。

3 生成式人工智能虚假信息传播风险评价指标体系构建

对生成式人工智能虚假信息传播风险的科学评价,需实现定性分析与定量分析的有机结合,而定量分析的有效开展,离不开一套系统、规范的风险评价指标体系作为基础支撑。结合拉斯韦尔信息传播理论与系统的文献梳理与分析,对生成式人工智能虚假信息传播的要素风险来源展开拆解,精准定位风险演化的关键节点,初步构建生成式人工智能虚假信息传播风险评价指标体系框架。随后,采用德尔菲法对指标进行科学筛选与优化,先后开展两轮专家意见征询。在经过两轮的专家意见征询后,删除逻辑重复矛盾和与实际获取数据困难、可操作性不强的指标。对评价指标进行明晰化、具体化处理,针对个别指标开展补充与修订,进一步提升其客观性与实操适配性。在专家组成员的专业指导与宝贵建议下,经过两次系统修改与完善,最终确立了包含 5 个准则层指标以及 27 个指标层指标的生成式人工智能虚假信息传播风险评价指标体系,见表 1。

4 生成式人工智能虚假信息传播风险评价指标体系分析

生成式人工智能虚假信息传播风险各维度之间存在相互转化、相互影响及动态反馈的复杂关系。借鉴层次分析法的核心框架与标准化流程,科学计算并确定评价指标体系中各项指标的权重。通过构建层次结构模型、构造判断矩阵、一致性检验等步骤,量化各风险指标的重要性程度,从而为风险评估提供客观、可操作的数据支持。

4.1 构建层次模型

通过已确立的有效评价指标,结合研究的核心

表 1　生成式人工智能虚假信息传播风险评价指标体系

目标层	准则层	指标层
生成式人工智能虚假信息传播风险评价指标体系 A_1	虚假信息传播主体风险 B_1	算法偏见程度 C_1
		技术工具的滥用频率 C_2
		动机行为的隐匿性 C_3
		对生成式 AI 的依赖程度 C_4
	虚假信息传播内容风险 B_2	虚假内容生成的频率 C_5
		AI 提示错误覆盖率 C_6
		信息来源模糊度 C_7
		热点场景贴合度 C_8
		内容迭代规避审核率 C_9
	虚假信息传播媒介风险 B_3	媒介覆盖规模 C_{10}
		信息扩散速率 C_{11}
		用户互动便捷性 C_{12}
		跨媒介内容迁移便捷性 C_{13}
		审核人员专业能力 C_{14}
		违规内容处置效率 C_{15}
	虚假信息传播受众风险 B_4	信息认知素养水平 C_{16}
		二次传播频率 C_{17}
		虚假信息参与度 C_{18}
		虚假信息关注度 C_{19}
		受众触达深度 C_{20}
	虚假信息传播效果风险 B_5	错误认知覆盖广度 C_{21}
		认知偏差固化深度 C_{22}
		认知纠正难度 C_{23}
		群体非理性行为强度 C_{24}
		公共资源消耗成本 C_{25}
		虚假信息舆论占比 C_{26}
		舆论极化程度 C_{27}

问题，按指标间内在相关性构建了自上而下的递进式层次结构模型，包含总目标层、准则层和指标层 3 个层级。总目标层聚焦生成式人工智能虚假信息传播风险综合评价，旨在全面评估该类风险的整体水平；准则层涵盖 5 项关键维度指标；指标层细化出 27 项具体指标，全面表征风险各环节及影响因素，为后续量化评价提供结构化支撑。

4.2　构建判断矩阵

在复杂系统的评价过程中，各指标间的关联逻辑往往具有非线性特征，难以通过简单加权平均实现科学表征，因此需借助系统化方法对指标进行综合分析与权重分配。于是在进行数据处理时，需要对每一层次结构的指标通过具体数值表现其重要性，也就是构造判断矩阵。判断矩阵是用于量化比较同一层级内不同因素重要性程度的标准化矩阵工具，其构建要求专家或评估者依据预设评价标准，对同一层级的指标进行两两成对比较。通过向 10 位专家发放问卷，邀请其依据 1-9 标度法对同一层级指标进行两两比较并量化赋值，为后续权重计算提供规范化的数值依据。10 位专家构成覆盖 AI、传播学、公共治理 3 大领域，高校学者 5 人，含 AI 算法、传播规律、数字共治等方向资深教授与副教授；社会从业者 5 人，主要为企业算法安全、行业合规、政策落地执行等一线实践专家。专家组合确保研究深度与实践适配性。准则层矩阵见表 2。

表 2　准则层矩阵

评价指标	虚假信息传播主体	虚假信息传播内容	虚假信息传播媒介	虚假信息传播受众	虚假信息传播效果
虚假信息传播主体	1	0.639 4	0.906 6	1.376 9	0.376 9
虚假信息传播内容	1.563 9	1	2.377 1	3.907 3	0.817 5
虚假信息传播媒介	1.103 1	0.420 7	1	2.377 1	0.827 2
虚假信息传播受众	0.726 2	0.255 9	0.420 7	1	0.733 9
虚假信息传播效果	2.653 1	1.223 3	1.209	1.362 5	1

4.3　一致性检验

一致性检验是为了证明上述判断矩阵符合正常逻辑规律。专家在对指标进行两两比较时可能会出现不一致的结论，所以有必要对已存的判断矩阵进行一致性检验以保证指标权重的合理性。同时学术界一般将 CR 作为判断矩阵一致性的标准，CR 为一致性指标 CI 和平均随机一致性指标 RI 的比值。若 CR＜0.1 则表明矩阵符合要求无须修改；否则，应请专家再次修正判断矩阵，最终使计算结果 CR＜0.1。一致性检验步骤：首先计算指标一致性 CI，$CI=(\lambda_{max}-n)/(n-1)$，$\lambda_{max}$ 为最大特征根，n 为判断矩阵阶数，CI 越大，说明判断矩阵越不一致；其次，计算一致性比例 CR，$CR=CI/RI$，当 CR＜0.1 时，判断矩阵的一致性符合要求。根据上述步骤，准则层指标权重见表 3。

由表 3 可知，CR＝0.049 1，小于 0.1，通过一致性检验，该层次分析法得到的权重结果是可靠的。生成式人工智能虚假信息传播风险评价准则层指标权重为：虚假信息传播主体0.146 4，虚假信息传播内容0.305 4，虚假信息传播媒介0.182 8，虚假信息传播受众0.105 1，虚假信息传播效果0.260 3。

表 3　准则层矩阵权重计算结果

评价指标	权重	λ_{max}	CI	CR
虚假信息传播主体	0.146 4	5.220 1	0.055	0.049 1＜0.1 一致性校验通过
虚假信息传播内容	0.305 4			
虚假信息传播媒介	0.182 8			
虚假信息传播受众	0.105 1			
虚假信息传播效果	0.260 3			

4.4 计算各指标权重

基于各个指标判断集成矩阵与权重的计算结果,一致性比率CR均小于0.1,均通过一致性检验,前后逻辑具备一致性。最终得到的生成式人工智能虚假信息传播风险评价指标体系的权重结果见表4。

5 对策与建议

在生成式人工智能风险快速演进的当下,传统单一主体治理模式已难以应对复杂的风险挑战,构建多方协同参与的治理体系成为迫切需求。因此基于协同治理理论,结合前文深度分析提炼的生成式人工智能虚假信息传播风险评价指标体系与权重,系统提出针对性风险治理策略。主要从政府、企业和社会3个维度的协同机制,构建政府主导统筹、企业执行落地、社会监督赋能的闭环联动体系,通过围绕虚假信息传播主体、虚假信息传播内容、虚假信息传播媒介、虚假信息传播受众、虚假信息传播效果这5大核心方面,打破单一治理的局限性,实现全方面精准防控。

(1)主体协同,构建虚假信息源头防控机制。生成式AI虚假信息传播的源头在于工具开发者、发布平台、个体传播者等多元主体,以政府、企业、社会三方联动为核心,实现虚假信息生成与传播的源头阻断。政府为统筹引领主体,核心是建立主体资质规范与责任划分体系,通过落实溯源备案、明确问责机制、共建信用档案与语料联盟,为源头管控提供制度与监管保障[37]。企业为直接执行主体,需将政府制度转化为技术措施与运营流程,通过嵌入数字水印、严格实名资质核验、算法监测识别恶意行为并快速管控,实现精准源头拦截。社会力量为补充监督主体,依托多元渠道举报违规主体线索,开展科普活动提升公众识别能力,引导公众规避高风险信息,从需求端压缩违规主体的生存空间。

(2)内容治理,建立精准审核与标准体系。虚假信息传播内容风险方面,政府作为统筹主体,核心是建立统一的判定标准与技术支撑体系,明确文本、图片等生成内容的水印标识规则,严禁篡改隐匿标识,通过跨平台标准互认破解漏判、过度管控问题。企业作为执行主体,将政府标准转化为落地流程,构建第三方审核与行业自律复合机制[38],对高风险主体常态化跟踪,对热点、公共事件类内容开启快速审核通道,以分层分级审核实现精准拦截。社会力量作为补充监督主体,畅通举报渠道鼓

表4 生成式人工智能虚假信息传播风险评价指标权重

目标层	准则层	权重	指标层	权重	综合权重
生成式人工智能虚假信息传播风险评价指标体系A_1	虚假信息传播主体风险B_1	0.146 4	算法偏见程度C_1	0.250 2	0.036 6
			技术工具的滥用频率C_2	0.244 0	0.035 7
			动机行为的隐匿性C_3	0.348 2	0.051 0
			对生成式AI的依赖程度C_4	0.157 6	0.023 0
	虚假信息传播内容风险B_2	0.305 4	虚假内容生成的频率C_5	0.163 0	0.049 8
			AI提示错误覆盖率C_6	0.127 0	0.038 8
			信息来源模糊度C_7	0.109 9	0.033 6
			热点场景贴合度C_8	0.383 3	0.117 1
			内容迭代规避审核率C_9	0.216 8	0.066 2
	虚假信息传播媒介风险B_3	0.182 8	媒介覆盖规模C_{10}	0.122 8	0.022 4
			信息扩散速率C_{11}	0.173 5	0.031 7
			用户互动便捷性C_{12}	0.063 8	0.011 7
			跨媒介内容迁移便捷性C_{13}	0.240 1	0.043 9
			审核人员专业能力C_{14}	0.102 3	0.018 7
			违规内容处置效率C_{15}	0.297 4	0.054 4
	虚假信息传播受众风险B_4	0.105 1	信息认知素养水平C_{16}	0.201 2	0.021 1
			二次传播频率C_{17}	0.152 1	0.016 0
			虚假信息参与度C_{18}	0.065 4	0.006 9
			主动搜索相关虚假信息的频率C_{19}	0.452 2	0.047 5
			受众触达深度C_{20}	0.129 1	0.013 6
	虚假信息传播效果风险B_5	0.260 3	错误认知覆盖广度C_{21}	0.205 2	0.053 4
			认知偏差固化深度C_{22}	0.132 5	0.034 5
			认知纠正难度C_{23}	0.063 3	0.016 5
			群体非理性行为强度C_{24}	0.133 3	0.034 7
			公共资源消耗成本C_{25}	0.099 9	0.026 0
			虚假信息舆论占比C_{26}	0.224 8	0.058 5
			舆论极化程度C_{27}	0.141 1	0.036 7

励公众举证，组建中立第三方评估团队定期核验平台甄别系统，反馈优化建议，形成协同治理闭环。

(3)媒介监管，阻断虚假信息扩散通道。生成式AI虚假信息传播媒介是风险扩散的核心载体，以政府、企业和其他社会力量三方联动为核心，通过明确责任、精准管控、监督反馈，在传播关键节点切断扩散通道。政府为统筹监管主体，核心是明确媒介平台的主体责任与法律义务，制定管控标准并构建常态化监管机制；重点规范算法推荐环节，通过法规约束平台不得利用算法放大虚假信息，降低高风险内容推荐权重。企业为直接执行主体，需将政府规则转化为平台落地细则，对违规信息和账号采取删除、限流等处置措施；针对公共卫生、金融政策等高风险话题内容强制增加人工复核，提高违规成本。社会力量为补充监督主体，依托多元广泛优势，通过多渠道举报算法滥用、虚假信息漏拦等问题，重点监督小众平台、私密社群等审核盲区，提升监督精准度与有效性。

(4)受众赋能，提升虚假信息辨别能力。虚假信息传播受众是风险的直接承受者，受众对生成式人工智能虚假信息的辨别能力直接决定风险危害程度。通过政府、企业、社会三方联动，建立分群体精准赋能、线上线下全覆盖、效果可评估的认知提升体系，从需求端增强受众风险抵御力，遏制虚假信息的传播动力。政府作为主导者与规划制定者，需提升自身技术甄别能力，出台专项赋能规划，联合教育机构构建全民终身数字素养课程体系，统筹资源保障偏远地区与特殊群体的赋能全覆盖[39]。企业作为场景载体与服务主体，依托平台设立辨别专区、设置高风险内容警示弹窗，同时通过优化模型训练、开发真伪检测工具，将政策规划落地为场景化服务。社会力量作为补充与监督主体，深入基层开展科普讲座、实操工作坊，助力老年人、低学历人群掌握辨别技巧[40]，同时收集受众需求反馈，推动赋能策略持续优化。

(5)构建修复体系，降低虚假信息传播危害。虚假信息传播效果直接关联个体权益、群体认知与社会信任，通过政府、企业、社会三方联动，最大程度降低虚假信息已造成的人身财产损失、舆论扰动、社会信任损耗等危害，并对无法完全消除的长期影响进行持续管理，实现风险危害的有效兜底。首先，政府为统筹主导主体，核心是整合资源搭建修复保障体系，既在舆情初期发布官方辟谣信息、启动监测引导机制，又协调多部门与跨区域资源，为受害者提供立案、法律援助等权益保障。其次，企业作为直接执行主体，依托平台数据优势定向推送辟谣内容、提供传播链路取证支持，并针对不同损害类型为用户提供科普纠偏、名誉修复、资金核查等差异化服务。最后，以社会力量为补充与监督主体，聚焦政府和企业的修复盲区，为受害者提供心理疏导、生活帮扶，协助数字技能薄弱群体对接维权渠道，降低维权门槛。

参考文献

[1] KHOO B, PHAN R C W, LIM C H. Deepfake attribution: on the source identification of artificially generated images[J]. Wiley Interdisciplinary Reviews:Data Mining and Knowledge Discovery, 2022, 12(3): e1438.

[2] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online [J]. Science, 2018, 359: 1146-1151.

[3] TORTORA L. Beyond discrimination: generative ai applications and ethical challenges in forensic psychiatry[J]. Frontiers in Psychiatry, 2024, 15: 1346059.

[4] ANTONIA C, TOM W, VINCENT D, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[5] 贾开，薛澜. 人工智能伦理问题与安全风险治理的全球比较与中国实践[J]. 公共管理评论，2021，3(1)：122-134.

[6] 周妍，沈天健. 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[J]. 编辑之友，2024(8)：75-83.

[7] LI B, QI P, LIU B, et al. Trustworthy AI: from principles to practices[J]. ACM Computing Surveys, 2023, 55(9): 1-46.

[8] 张璐. 通用人工智能风险治理与监管初探：ChatGPT引发的问题与挑战[J]. 电子政务，2023，20(9)：14-24.

[9] 郑满宁. 人工智能技术下的新闻业：嬗变、转向与应对：基于ChatGPT带来的新思考[J]. 中国编辑，2023(4)：35-40.

[10] 张凌寒. 搜索引擎自动补足算法的损害及规制[J]. 华东政法大学学报，2019，22(6)：31-45. 10.

[11] CHAMBIAL S, BUDHIA R, PANDEY T, et al. Deep fake generation and detection[J]. Advances in Distributed Computing and Machine Learning, 2023, 660(4): 533-543.

[12] 冯由玲，康鑫，周金娉，等. 基于BERT-BiLSTM混合模型的社交媒体虚假信息识别研究[J]. 情报科学，2024，42(6)：89-98.

[13] 何宇华，李霞. 生成式人工智能虚假信息治理的新挑战及应对策略：基于敏捷治理的视角[J]. 治理研究，2024，40(4)：142-156，160.

[14] TAMBINI D. A framework for media freedom [J]. Media, Culture & Society, 2020, 42(4): 578-594.

[15] 魏屹东，李煜慧. 生成式人工智能引发的意识形态风险及其应对[J]. 理论探索，2025(2)：17-25.
[16] 王延川，赵靖. 生成式人工智能诱发意识形态风险的逻辑机理及其应对策略[J]. 河南师范大学学报(哲学社会科学版)，2023，50(4)：1-7.
[17] 蒲清平，向往. 生成式人工智能：ChatGPT 的变革影响、风险挑战及应对策略[J]. 重庆大学学报(社会科学版)，2023，29(3)：102-114.
[18] 刘祖兵. 论 Sora 的伦理风险与我国治理因应：也谈我国参与人工智能算法伦理全球治理的基本路径[J]. 河海大学学报(哲学社会科学版)，2024，26(5)：99-112.
[19] 贾开，蒋余浩. 人工智能治理的三个基本问题：技术逻辑、风险挑战与公共政策选择[J]. 中国行政管理，2017(10)：40-45.
[20] 贺小石. 人工智能赋能教育的信息传播风险与治理路径[J]. 编辑学刊，2025(3)：25-29.
[21] 周毅. 试论网络信息内容治理主体构成及其行动转型[J]. 电子政务，2020(12)：41-51.
[22] 林雨佳. 生成式人工智能对信息治理的挑战与应对[J]. 苏州大学学报(哲学社会科学版)，2025，46(2)：104-115.
[23] 朱嘉珺. 生成式人工智能虚假有害信息规制的挑战与应对：以 ChatGPT 的应用为引[J]. 比较法研究，2023(5)：34-54.
[24] 付安玲，张书铭. 生成式人工智能背景下网络意识形态风险治理的逻辑理路与治理实践[J]. 学习论坛，2025(5)：5-13.
[25] 邓宏光，王雪璠. 生成式人工智能虚假信息治理的风险与应对[J]. 理论月刊，2024(9)：115-129.
[26] 郑林，章荣君. 人工智能赋能主流意识形态传播的风险及其规制[J]. 理论探索，2025(4)：54-62.
[27] 漆晨航. 生成式人工智能的虚假信息风险特征及其治理路径[J]. 情报理论与实践，2024，47(3)：112-120.
[28] 杜智涛，张伊晴. 平台的人工智能披露义务与网络虚假信息治理责任[J]. 新媒体与社会，2025(2)：71-87，462.
[29] 常琳，叶凤云. 生成式人工智能用户虚假信息规避行为研究[J]. 情报理论与实践，2025，48(2)：102-112.
[30] 杨晓彤. 生成式人工智能的意识形态隐喻及其具象化风险应对[J]. 马克思主义研究，2025(9)：133-144.
[31] 王俊秀. ChatGPT 与人工智能时代：突破、风险与治理[J]. 东北师大学报(哲学社会科学版)，2023(4)：19-28.
[32] 柳亦博，王玉雪. 走向生命行政：生成式人工智能参与行政决策的伦理边界与行动原则[J]. 中共天津市委党校学报，2025，27(5)：32-42.
[33] 赵浩宇，韩正彪，殷淑琪. 信息实践中的“变通”艺术：生成式人工智能用户应对限制的战术研究[J]. 情报资料工作，2026，47(1)：103-111.
[34] 孙鲲鹏，王丹，肖星. 互联网信息环境整治与社交媒体的公司治理作用[J]. 管理世界，2020，36(7)：106-132.
[35] 李明德，朱妍. 复杂舆论场景中信息内容传播风险研究[J]. 情报杂志，2021，40(12)：112-119.
[36] 王国华，肖林，汪娟，等. 论舆论场及其分化问题[J]. 情报杂志，2012，31(8)：1-4，20.
[37] 苏宇. 大型语言模型生成虚假信息风险的法律治理[J]. 环球法律评论，2025，47(6)：36-52.
[38] 支振锋. 生成式人工智能大模型的信息内容治理[J]. 政法论坛，2023，41(4)：34-48.
[39] 李白杨，唐昆. AIGC 背景下全民数字素养教育的内涵变革与应对策略[J]. 图书与情报，2024(3)：32-39.
[40] 付少雄，曾源来，孙岚，等. 组态视角下短视频虚假健康信息传播行为的影响因素研究[J]. 图书馆杂志，2024，43(12)：103-116.

Construction of Evaluation Index System for Dissemination Risks of Generative Artificial Intelligence Disinformation

WEI Changchang，LI Chan

(School of Public Administration and Law，Fujian Agriculture and Forestry University，Fuzhou 350002，China)

Abstract：While generative artificial intelligence technology improves the efficiency of information production，it also exacerbates the creation and spread of false information，leading to social instability issues. Based on risk management theory，the current governance situation of false information dissemination in the context of generative artificial intelligence was systematically reviewed，and its dissemination characteristics and social impacts were analyzed. Combining the core framework of information dissemination theory，a risk assessment indicator system for the dissemination of false information in generative artificial intelligence was constructed from five dimensions：source of dissemination，content of dissemination，dissemination medium，audience，and dissemination effect. The analytic hierarchy process was used to determine the weight of each level of indicators. Ultimately，a complete evaluation system containing 27 specific indicators was developed. It is found that content dissemination risk has the highest weight in the indicator system，followed by dissemination effect risk，dissemination medium risk，and others. The research results can provide the useful reference and guidance for accurately identifying and scientifically assessing the risks of false information dissemination through generative artificial intelligence.

Keywords：dissemination risks of disinformation；generative artificial intelligence；evaluation index system