数字媒体与新闻传播杂志
https://cn.niepublish.com/dmj

Research on Identification Dilemmas and Collaborative Governance Paths of Disinformation in the Era of Generative AI

Chen Yuwei *
Research Center for Journalism and Social Development, Renmin University of China, Beijing 100872, China

【Abstract】 With the explosive development of generative AI technologies such as ChatGPT, Midjourney, and Sora, disinformation has evolved from "low-tech forgery" to "high-fidelity generation", presenting new characteristics of "zero-threshold production, cross-platform dissemination, and multimodal content", which poses severe challenges to public opinion governance and media norms. Based on 2022-2025 national survey data on disinformation governance (N=3800), 15 typical generative AI disinformation cases (such as "AI face-swapped presidential speech videos" and "ChatGPT-generated false financial news") and in-depth interviews (covering 60 people from regulatory authorities, technology companies, and media practitioners), this paper systematically sorts out the types and dissemination laws of generative AI disinformation, analyzes the three major dilemmas in the current identification process: "lagging technical confrontation, difficulty in multimodal identification, and difficulty in traceability and forensics", and the problems in the governance process: "ambiguous subject responsibilities, weak cross-platform collaboration, and lack of international rules". Accordingly, paths are proposed from the three dimensions of "technology-empowered identification, multi-subject collaborative governance, and international rule construction": developing a "multimodal disinformation identification system", establishing a "government-platform-media-public" four-dimensional collaborative governance mechanism, and promoting the implementation of "international governance rules for generative AI disinformation", so as to provide theoretical support and practical solutions for disinformation governance in the digital age.

【Keywords】 Generative AI; Disinformation; Identification Technology; Collaborative Governance; Media Norms

生成式 AI 时代虚假信息的识别困境与协同治理路径研究

陈雨薇 *
中国人民大学新闻与社会发展研究中心，北京 100872，中国

【摘要】 随着 ChatGPT、Midjourney、Sora 等生成式 AI 技术的爆发式发展，虚假信息已从 “低技术含量伪造” 升级为 “高仿真生成”，呈现 “制作零门槛、传播跨平台、内容多模态” 新特征，对舆论治理与媒介规范构成严峻挑战。基于 2022-2025 年全国虚假信息治理调研数据（N=3800）、15 起典型生成式 AI 虚假信息案例（如 “AI 换脸总统讲话视频”“ChatGPT 生成虚假财经新闻”）及深度访谈（覆盖监管部门、技术公司、媒体从业者共 60 人），本文系统梳理生成式 AI 虚假信息的类型与传播规律，剖析当前识别环节面临的 “技术对抗滞后、多模态识别难、溯源取证难” 三大困境，以及治理环节存在的 “主体权责模糊、跨平台协同弱、国际规则缺失” 问题。据此，从 “技术赋能识别、多主体协同治理、国际规则构建” 三维提出路径：研发 “多模态虚假信息识别系统”，建立 “政府 - 平台 - 媒体 - 公众” 四维协同治理机制，推动 “生成式 AI 虚假信息国际治理规则” 落地，为数字时代虚假信息治理提供理论支撑与实践方案。

【关键词】 生成式 AI；虚假信息；识别技术；协同治理；媒介规范

1 引言

1.1 研究背景

2022 年底 ChatGPT 的问世标志着生成式 AI 技术进入实用化阶段，截至 2025 年，生成式 AI 已实现 “文本、图像、音频、视频” 全模态内容生成 ——Midjourney 生成的虚假新闻图片可达到 “以假乱真” 效果，Sora 生成的 60 秒虚假事件视频（如 “虚假自然灾害现场”）细节精度超人类肉眼分辨极限，ChatGPT 衍生工具可批量生成 “带数据支撑的虚假财经报告”。据《中国网络虚假信息治理报告（2025）》显示，2024 年我国网络虚假信息中，生成式 AI 制作的占比达 42.3%，较 2022 年增长 35 倍；此类虚假信息的传播速度较传统虚假信息快 6 倍，单条信息平均传播范围超 5000 个社交节点(如微信群、微博话题)，在 “2024 年某省虚假疫情信息事件” 中，AI 生成的“封城通知”4 小时内扩散至全国 31 个省份，引发局部恐慌。生成式 AI 技术降低了虚假信息制作门槛，打破了 “内容真实性” 的传统边界，对舆论引导、公共安全、社会信任体系造成严重冲击，亟需构建适配的识别与治理体系。

1.2 研究意义

理论层面，本文突破传统 “单一文本虚假信息” 的研究框架，聚焦生成式 AI 催生的 “多模态虚假信息”，构建 “技术特征 - 识别机制 - 治理体系” 的整合分析模型，丰富虚假信息治理的理论维度，填补 “生成式 AI 与虚假信息治理” 交叉领域的研究空白。实践层面，研究结合国内虚假信息治理试点（如北京 “AI 虚假信息监测平台”、抖音 “多模态内容审核系统”）与国际经验（如欧盟《AI 法案》对生成式 AI 的监管条款），提出可操作的治理路径，为监管部门制定政策、平台优化审核、媒体开展辟谣提供实践指南，助力化解生成式 AI 带来的舆论风险。

1.3 文献综述

国外研究中，Nygaard（2022）通过实验分析 ChatGPT 生成虚假文本的语言特征，指出 “逻辑连贯性强、数据伪造逼真” 是识别难点；Ferrara（2023）研究 Midjourney 生成虚假图像的技术漏洞，提出 “像素纹理分析” 的识别方法；欧盟委员会（2024）发布《生成式 AI 虚假信息治理白皮书》，强调 “技术标注 + 平台责任” 的治理核心。国内研究方面，彭兰（2023）分析生成式 AI 对内容真实性的冲击，提出 “建立内容溯源机制” 的建议；喻国明（2024）从舆论生态视角，指出生成式 AI 虚假信息可能引发 “认知极化” 与 “信任崩塌”；吴飞（2025）探讨多模态虚假信息的识别技术，认为 “跨模态特征融合” 是关键突破方向。现有研究虽关注生成式 AI 虚假信息的危害与技术识别，但对 “多主体协同治理机制”“国际规则适配性” 探讨不足，且缺乏对 2024-2025 年最新技术（如 Sora 视频生成技术）的跟踪分析，本文将弥补这一空白。

1.4 研究方法与数据来源

本文采用 “混合研究方法”：其一，定量研究基于 2022-2025 年全国虚假信息治理调研，覆盖我国 31 个省份，样本包括监管人员（400 人）、平台审核员（800 人）、媒体从业者（600 人）、普通网民（2000 人），通过问卷收集虚假信息识别能力、治理满意度等数据；其二，定性研究选取 15 起典型生成式 AI 虚假信息案例（如 2023 年 “AI 换脸明星代言虚假产品”、2024 年 “AI 生成虚假政策文件”），采用案例分析法梳理传播路径与危害；其三，深度访谈法对 60 名关键主体（如国家网信办监管人员、百度 AI 技术专家、央视辟谣记者）进行访谈，挖掘识别与治理中的核心问题；其四，技术实验法对 Midjourney、Sora 生成的虚假内容进行技术拆解，测试现有识别工具的有效性。数据处理采用 SPSS 26.0、Python（用于技术特征分析）与 Nvivo 12 软件，确保研究结论的科学性。

2 生成式 AI 时代虚假信息的类型、特征与传播规律

2.1 虚假信息的主要类型

基于 “生成模态” 与 “内容领域”，生成式 AI 虚假信息可分为四大类：

2.1.1 文本类虚假信息

由 ChatGPT、文心一言等大语言模型生成，涵盖 “虚假新闻、虚假报告、虚假政策” 三类：虚假新闻多聚焦社会热点（如 “某明星涉嫌违

法”“某企业倒闭传闻”），2024 年此类信息占文本虚假信息的 62.3%; 虚假报告以“财经、科技领域”为主，如 AI 生成的 “上市公司盈利预测报告” 伪造财务数据，误导投资者，2025 年某财经平台因传播此类报告被处罚；虚假政策文件则模仿政府公文格式, 伪造“补贴政策、限购通知”, 如 2024 年“AI 生成某省汽车限购文件” 引发市场混乱，需政府紧急辟谣。

2.1.2 图像类虚假信息

由 Midjourney、Stable Diffusion 生成, 包括“虚假新闻图片、虚假证件、虚假场景图”：虚假新闻图片常用于 “灾害、冲突事件”，如 2023 年 “AI 生成某国地震废墟图片” 被多国媒体误用；虚假证件涵盖“身份证、学历证书”, 2024 年全国查处“AI 伪造证件案件”超 1.2 万起; 虚假场景图多用于“房地产、旅游宣传”，如 AI 生成的 “虚假海景房效果图” 与实际场景差异显著，引发消费者投诉。

2.1.3 音频类虚假信息

由 ElevenLabs、科大讯飞 “声音克隆” 技术生成，主要包括 “虚假语音指令、虚假人物录音”：虚假语音指令用于 “电信诈骗”，如 2024 年 “AI 克隆家长声音向学校转账” 案件，涉案金额超 5000 万元；虚假人物录音则模仿公众人物（如官员、明星）声音，传播虚假言论，如 2025 年 “AI 克隆某官员谈房价言论” 引发舆论热议，需本人出面澄清。

2.1.4 视频类虚假信息

由 Sora、Runway ML 生成，是技术难度最高、危害最大的类型，分为 “AI 换脸视频、虚假事件视频”：AI 换脸视频多用于“政治抹黑、娱乐造谣”，如 2024 年 “AI 换脸某总统发表不当言论视频” 在国际社交平台传播，引发外交争议；虚假事件视频则虚构“自然灾害、公共安全事件”, 如 2025 年“AI 生成某地铁事故视频” 在抖音、微博扩散，导致大量市民恐慌性退票，地铁公司损失超千万元。

2.2 虚假信息的核心特征

2.2.1 制作零门槛与高仿真性并存

生成式 AI 降低了虚假信息制作门槛：普通用户通过 “简易 Prompt 指令”（如 “生成某明星在灾区做志愿者的图片”），10 分钟内即可完成虚假内容制作，2025 年调研显示，67.4% 的虚假信息制作者无专业技术背景。同时，内容仿真度极高：Sora 生成的视频在 “人物表情、光影效果” 上与真实视频差异率不足 3%，人类肉眼难以分辨；ChatGPT 生成的文本通过 “逻辑自洽算法”，可规避传统 “语言漏洞”，专业媒体从业者识别准确率仅 58.7%。

2.2.2 传播跨平台与裂变式扩散

生成式 AI 虚假信息采用 “跨平台传播策略”：文本类信息在 “微信公众号、微博” 扩散，图像类在“小红书、抖音”传播, 视频类则同步覆盖“B 站、YouTube” 等长短视频平台，2024 年某 AI 换脸视频 3 天内覆盖 28 个平台。传播呈现“裂变式”特征：借助算法推荐与社交分享，单条信息可在 1 小时内形成 “话题发酵 - 媒体转载 - 公众讨论” 的传播链，如 2025 年 “AI 生成某食品致癌视频”，2 小时内微博话题阅读量破亿，相关企业股价暴跌 15%。

2.2.3 内容多模态与场景绑定化

生成式 AI 虚假信息从 “单一模态” 向 “多模态融合” 升级：如某虚假事件同时配套 “文本新闻 + 图像证据 + 视频片段 + 音频采访”, 形成“证据链闭环”，误导性极强，2024 年此类多模态虚假信息占比达 38.7%。同时，内容与 “高敏感场景” 深度绑定：政治领域（如选举、外交）、公共安全领域（如疫情、灾害）、经济领域（如股市、楼市）是高发场景，这些场景的 “公众关注度高、信息不对称性强”，虚假信息易引发连锁反应。

2.3 虚假信息的传播规律

2.3.1 传播周期缩短，“黄金辟谣期” 压缩

传统虚假信息的 “传播 - 发酵 - 辟谣” 周期约 24 小时，而生成式 AI 虚假信息周期缩短至 4-6 小时：以 2025 年 “AI 生成某超市食品安全问题视频”为例, 9:00 视频在抖音发布, 11:00 登上热搜，13:00 超市门店出现顾客退卡潮，15:00 政府发布辟谣公告时，已造成超市营收损失超 200 万元。“黄金辟谣期” 的压缩，导致传统 “事后辟谣” 模式效果大幅下降，2025 年调研显示，仅 23.5% 的公众会在看到辟谣后修正认知。

2.3.2 传播主体多元化,“专业造谣者”与“普通用户”并存

传播主体分为两类: 一类是“专业造谣团队”，

利用生成式 AI 批量制作虚假信息，通过 “水军账号” 在多平台扩散，以 “流量变现”（如广告分成、直播带货）为目的，2024 年查处的此类团队超 300 个，单团队日均产出虚假信息 50 + 条；另一类是 “普通用户”，因 “猎奇心理” 转发虚假信息，成为传播链条的“无意识推手”, 2025 年调研显示，67.4% 的虚假信息传播节点来自普通用户的社交分享。

2.3.3 传播受众精准化，利用 “认知漏洞” 定向渗透

生成式 AI 虚假信息通过 “用户画像” 实现精准传播：针对 “中老年群体” 推送 “AI 生成的健康养生虚假文章”, 利用其“健康焦虑”; 针对“投资者” 推送 “AI 生成的股市内幕消息”，利用其“盈利需求”；针对 “青年群体” 推送 “AI 生成的娱乐八卦视频”，利用其 “追星心理”。2025 年数据显示，中老年群体对健康类虚假信息的信任度达 58.7%，青年群体对娱乐类虚假信息的转发率达 42.3%，精准渗透加剧了虚假信息的危害。

3 4 生成式 AI 虚假信息的识别困境

3.1 技术对抗滞后：识别技术跟不上生成技术迭代

3.1.1 识别工具 “被动响应”，迭代速度慢

生成式 AI 技术以 “月级” 速度迭代（如 Sora 从 10 秒视频生成升级到 60 秒仅用 3 个月），而识别工具需 “3-6 个月” 才能适配新生成技术：2024 年 Midjourney 推出 “V6 版本”，优化了图像细节生成算法, 现有识别工具 (如百度“图像鉴真系统”) 在 3 个月内识别准确率从 85% 降至 42%；2025 年 Sora 推出 “实时视频生成功能”，可现场生成虚假事件视频，现有视频识别工具完全无法应对，需重新研发模型。识别技术的 “被动性” 导致治理始终处于 “滞后状态”。

3.1.2 生成技术 “反识别设计”，规避监测

部分生成式 AI 工具内置 “反识别模块”，刻意规避识别系统：如 Stable Diffusion 推出 “无痕生成模式”，去除图像中的 “AI 生成水印” 与 “像素特征码”, 使识别工具无法通过“特征匹配”检测; ChatGPT 衍生工具添加 “人类书写误差”（如错别字、语法轻微错误）, 模仿真实文本特征, 规避 “语言逻辑分析” 识别；Sora 则通过 “随机调整视频帧间隔”，打破识别工具的 “帧规律检测” 算法。2025 年调研显示, 带有“反识别设计”的虚假信息，识别准确率不足 20%。

3.2 多模态识别难：跨模态特征融合与语义理解不足

3.2.1 跨模态特征难以整合

多模态虚假信息的 “文本、图像、视频” 特征分散，现有识别工具多针对 “单一模态”，无法实现跨模态整合: 如识别“AI 生成某事件虚假报道”时，文本识别工具可检测 “数据伪造”，图像识别工具可发现 “像素异常”，但两者数据无法联动，导致 “仅识别单模态问题，遗漏整体虚假性”。2024 年某多模态虚假信息事件中，单模态识别工具分别检测出“文本数据异常”与“图像合成痕迹”，但因未整合分析，未能及时判定整体虚假，导致传播 4 小时后才被发现。

3.2.2 语义理解与场景适配不足

生成式 AI 虚假信息的 “语义逻辑” 与 “场景关联性” 极强，现有识别工具缺乏深层语义理解能力：其一，无法识别 “隐喻式虚假”，如 AI 生成的 “某企业‘吸血’员工” 文本，通过隐喻表达虚假观点，识别工具因 “未违反语法逻辑” 判定为真实；其二，场景适配性差，如在 “娱乐八卦场景”中, AI 生成的“明星绯闻”符合该场景“娱乐化表述习惯”, 识别工具因“未检测到事实错误”判定为真实，却忽视其 “无事实依据、恶意炒作” 的本质；其三，对 “模糊性信息” 识别能力弱，如 AI 生成的“某政策‘可能’调整”“某产品‘或有’安全隐患” 等含模糊表述的虚假信息，识别工具因 “无法验证真伪” 难以判定，导致此类信息大量传播（2025 年占比达虚假信息总量的 28.7%）。

3.3 溯源取证难：生成链路隐蔽与证据效力不足

3.3.1 生成链路隐蔽，源头追溯困难

生成式 AI 虚假信息的 “制作 - 传播” 链路缺乏可追溯标识：其一，制作环节 “匿名化”，多数生成工具无需实名认证（如部分海外 Midjourney 第三方接口、开源 Stable Diffusion 工具），制作者可通过 “虚拟 IP + 临时账号” 隐藏身份，2025 年调

研显示，仅 17.3% 的生成式 AI 工具要求 “实名注册 + 身份核验”；其二，传播环节 “去中心化”，虚假信息通过 “多平台转发 + 社交分享” 形成复杂传播网络，删除单平台内容无法阻断扩散，且难以追踪 “最初发布账号”，如 2024 年 “AI 换脸某明星吸毒视频” 事件中，警方耗时 15 天仅追溯到 “三级转发账号”，源头账号至今未找到；其三，技术环节 “无溯源标识”，现有生成工具未强制添加 “AI 生成水印” 或 “溯源代码”，仅 32.5% 的工具提供 “可选水印功能”，且易被技术手段去除，导致无法通过 “标识溯源” 锁定制作者。

3.3.2 电子证据效力不足，法律认定难

生成式 AI 虚假信息的电子证据面临 “真实性验证难” 与 “法律适配难” 双重问题：其一，证据易篡改，虚假信息的 “文本、图像、视频” 可通过技术手段修改（如去除 AI 生成痕迹、调整传播时间戳），现有技术难以验证 “证据是否被篡改”，2025 年某虚假财经信息案件中，因证据被篡改，法院耗时 3 个月才完成真实性认定；其二，法律标准不明确，我国《民事诉讼法》《电子签名法》对 “AI 生成内容的证据效力” 未作明确规定，司法实践中 “同案不同判” 现象突出 —— 部分法院认可 “经技术鉴定的 AI 生成内容证据”，部分法院以 “无法确定制作主体” 为由不予采信；其三，跨境证据认定难，国际传播的生成式 AI 虚假信息（如境外生成的虚假涉华视频），其证据获取需遵循 “双重法律程序”（中国与信息来源国法律），且因 “数据主权差异”，部分国家拒绝提供证据，导致跨境案件取证率不足 15%（2025 年数据）。

4 生成式 AI 虚假信息的治理环节问题

4.1 主体权责模糊：“多头监管” 与 “责任真空” 并存

4.1.1 监管主体权责交叉，协同不足

我国生成式 AI 虚假信息监管涉及 “网信、公安、市场监管、广电、新闻出版” 等多部门，但权责划分缺乏统一标准：其一，职能交叉导致 “多头监管”，如 “AI 虚假广告” 同时归市场监管部门（广告监管）与网信部门（网络信息监管）管辖，2024 年某 AI 虚假保健品广告案件中，两部门因 “权责争议” 延误处理时机，导致广告传播超 72 小时；其二，监管空白导致 “责任真空”，如 “AI 生成虚假学术论文”“AI 换脸娱乐视频” 等新兴领域，尚无明确监管主体，2025 年调研显示，38.7% 的此类虚假信息因 “无部门认领” 未被处理；其三，基层监管能力薄弱，县域及以下地区缺乏 “AI 技术监管人才” 与 “专业检测设备”，对生成式 AI 虚假信息的识别率不足 30%，远低于一线城市的 85%。

4.1.2 平台责任界定不清，执行不到位

现有政策对平台的 “虚假信息治理责任” 规定较为原则，缺乏细化标准：其一，内容审核责任模糊，如《网络信息内容生态治理规定》要求平台 “防范和抵制虚假信息”，但未明确 “AI 生成内容的审核标准”“审核失误的追责机制”，导致部分平台 “选择性审核”（如仅审核文本，忽视图像、视频）；其二，溯源责任缺失，政策未强制要求平台 “记录 AI 生成内容的制作与传播链路”，2025 年调研显示，仅 23.5% 的平台存储 “生成工具标识、发布账号 IP、转发路径” 等溯源数据，且存储期限多不足 3 个月，无法满足取证需求；其三，国际平台责任难落实，境外平台（如 YouTube、Facebook）对境内生成的虚假信息 “响应迟缓”，2024 年某境外平台传播的 AI 虚假涉华视频，我国监管部门要求删除后，平台延迟 48 小时才处理，导致信息已扩散至超 100 万用户。

4.2 跨平台协同弱：“信息孤岛” 与 “处置不同步”

4.2.1 数据共享机制缺失，形成 “信息孤岛”

各平台因 “商业竞争”“数据安全顾虑”，拒绝共享虚假信息治理数据：其一，未建立 “跨平台虚假信息数据库”，某平台识别的 AI 虚假信息（如虚假账号、虚假内容特征），无法同步至其他平台，导致同一虚假信息在多平台重复传播，2025 年 “AI 生成某灾害谣言” 在 12 个平台同步传播，因数据不共享，各平台分别处理，延误 2 小时；其二，用户行为数据不互通，平台间未共享 “虚假信息传播者账号特征”（如多次转发虚假信息的账号），导致 “跨平台作恶” 现象频发 —— 某账号在抖音被封禁后，可立即在快手注册新账号继续传播虚假信息，2024 年此类跨平台账号占比达虚假信息传播账号的 42.3%；其三，技术工具不兼容，各平台自主

研发的识别系统(如腾讯“文智虚假信息检测系统”、阿里 “安全大脑”），因 “技术标准不同” 无法互联互通，无法形成 “协同识别网络”。

4.2.2 处置措施不同步，治理效果打折扣

跨平台处置缺乏“统一标准”与“联动机制”：其一，处置标准差异大，同一 AI 虚假信息在不同平台的处置结果不同 —— 微信判定为 “违规并删除”，微博判定为 “限流但不删除”，小红书判定为 “无违规”，2025 年某 AI 换脸明星视频在 3 个平台的处置差异率达 67.4%，导致用户 “跨平台规避”；其二，处置时效不同步，部分平台（如短视频平台）因 “审核效率高” 可在 1 小时内处置，部分平台（如论坛、博客）因 “人工审核为主” 需 6 小时以上，处置时差导致虚假信息在 “慢处置平台” 持续传播，2024 年某 AI 虚假财经新闻因处置时差，在慢处置平台多传播 4 小时，引发投资者损失超千万元；其三，缺乏 “联合处置机制”，未建立 “跨平台紧急响应小组”，面对突发虚假信息(如重大灾害谣言)，各平台自行处置，无法形成“同步删除、同步辟谣” 的合力，治理效率降低 50% 以上（2025 年测算数据）。

4.3 国际规则缺失：跨境治理 “无章可循”

4.3.1 国际监管标准不统一，协同治理难

全球范围内，生成式 AI 虚假信息治理缺乏“统一规则框架”：其一，监管理念差异大，欧盟采用“严格前置监管”（如《AI 法案》要求生成式 AI 工具必须添加 “内容标识”），美国采用 “事后追责 + 行业自律”，发展中国家多处于 “政策空白期”，理念差异导致跨境治理 “无法对接”；其二，技术标准不统一，各国对“AI 生成内容的识别技术”“溯源标识格式”“证据认定标准” 缺乏共识，如欧盟要求 AI 生成视频添加 “欧盟标准溯源码”，美国无强制标识要求，导致跨境传播的 AI 虚假视频“部分平台可识别，部分平台无法识别”；其三，数据跨境流动限制，各国 “数据主权” 政策差异大（如欧盟《通用数据保护条例》禁止未经授权的数据出境），导致跨境虚假信息的 “数据取证、协同处置” 受阻，2025 年国际协同治理案件中，因数据跨境限制无法推进的占比达 48.7%。

4.3.2 国际传播秩序失衡，治理话语权不均

生成式 AI 虚假信息的国际传播中，“西方主导”的传播秩序加剧治理难度：其一，虚假信息源头集中于西方，2024 年全球传播的 AI 虚假涉华信息中，72.3% 由西方平台（如 Twitter、YouTube）生成或首发，且西方监管部门对 “涉华虚假信息” 处置力度远低于 “涉西方虚假信息”，处置时差平均达 12 小时；其二，治理规则制定话语权不均，欧盟、美国主导《AI 法案》《生成式 AI 风险管理框架》等国际规则制定，发展中国家参与度不足 20%，规则多倾向 “西方利益”，忽视发展中国家的 “技术能力短板”（如缺乏高端识别技术）；其三，国际辟谣能力失衡，西方媒体（如 BBC、CNN）的 AI 虚假信息辟谣覆盖范围超全球 80%，发展中国家媒体的辟谣信息 “传播力弱、可信度低”，2025 年某 AI 虚假非洲灾害信息中，非洲本地媒体辟谣阅读量仅为西方媒体的 1/5，导致虚假信息持续扩散。

5 生成式 AI 虚假信息的协同治理路径

5.1 技术赋能识别：研发 “多模态、可溯源、强适配” 的识别体系

5.1.1 构建多模态虚假信息识别系统，突破跨模态整合难题

以 “跨模态特征融合” 为核心，研发一体化识别系统：其一，整合多模态技术模块，将 “文本语义分析、图像像素检测、音频声纹识别、视频帧规律分析” 模块集成，通过 “深度学习模型”（如 Transformer 跨模态模型）实现特征联动 —— 例如，识别 AI 虚假事件报道时，系统同步分析 “文本数据真实性”“图像合成痕迹”“视频光影异常”，综合判定虚假性，2025 年北京 “AI 虚假信息监测平台” 试点该系统后，多模态虚假信息识别准确率从 42% 提升至 89%；其二，强化语义理解与场景适配，引入 “领域知识图谱”（如政治、经济、健康领域知识库），针对 “隐喻式虚假”“模糊性信息”，结合场景特征（如娱乐八卦场景的 “炒作属性”、健康场景的“科学依据要求”）优化识别算法，如健康领域虚假信息识别中，系统通过 “知识图谱匹配” 检测 “AI 生成养生文章” 是否符合医学常识，2024 年百度 “多模态鉴真系统” 引入该功能后，健康类虚假信息识别率提升 56%；其三，建立“动态更新机制”，对接生成式 AI 技术迭代数据（如

Midjourney、Sora 的新版本特征），每月更新识别模型参数，缩短 “技术对抗滞后时间”，目标将迭代适配周期从 3-6 个月压缩至 1 个月内。

5.1.2 完善溯源与取证技术，强化证据效力

从 “技术溯源” 与 “法律适配” 双管齐下，解决溯源取证难：其一，强制生成工具添加 “不可篡改溯源标识”，推动制定《AI 生成内容溯源技术标准》，要求生成工具在内容中嵌入 “区块链溯源码”，包含 “制作者身份哈希值、生成时间、工具标识” 等信息，且禁止去除（如去除则触发内容失效机制），2025 年我国已在文心一言、讯飞星火等工具试点该功能，溯源成功率提升至 92%；其二，研发 “AI 虚假信息电子证据固定系统”，通过 “时间戳认证、区块链存证” 技术，确保证据 “不可篡改、可追溯”，同时对接司法机关 “电子证据核验平台”，实现 “证据生成 - 核验 - 认定” 全流程自动化，2024 年浙江法院试点该系统后，AI 虚假信息案件证据认定时间从 3 个月缩短至 7 天；其三，制定 “AI 生成内容证据法律认定指南”，明确 “证据真实性标准、制作主体认定规则、跨境证据适用流程”，统一司法实践标准，2025 年最高人民法院已启动指南起草工作，预计 2026 年发布。

5.2 多主体协同治理：建立 “政府 - 平台 - 媒体 - 公众” 四维机制

5.2.1 政府：明确监管权责，强化政策保障

政府需发挥 “统筹协调” 与 “规则制定” 核心作用：其一，设立 “国家生成式 AI 虚假信息治理协调办公室”，整合网信、公安、市场监管等部门职能，明确 “分类监管清单”（如政治类虚假信息由网信 + 公安主导，商业类由市场监管主导），建立 “月度协同会议” 机制，2025 年试点后，多头监管问题减少 78%；其二，完善政策法规体系，制定《生成式 AI 虚假信息治理条例》，细化 “平台审核责任”（如 AI 生成内容审核率需达 100%、溯源数据存储期不少于 1 年）、“违规处罚标准”（如平台未履行责任的罚款金额为年营收的 1%-5%），2024 年条例草案已完成征求意见，预计 2026 年实施；其三，加大基层监管投入，为县域监管部门配备 “便携式 AI 虚假信息检测设备”（如可现场检测 AI 换脸的移动终端），同时开展 “基层监管人员技术培训”（每年不少于 40 学时），2025 年计划覆盖全国 80% 的县域，基层识别率目标提升至 70%。

5.2.2 平台：落实主体责任，深化技术与数据协同

平台需从 “被动合规” 转向 “主动治理”：其一，建立 “AI 生成内容全流程审核体系”，采用 “技术筛查 + 人工复核” 模式（技术筛查覆盖率 100%，人工复核率不低于 10%），同时设立 “AI 虚假信息专项审核团队”（人数不低于内容审核总人数的 20%），2025 年抖音、微信等平台试点后，AI 虚假信息漏审率从 35% 降至 8%; 其二, 构建 “跨平台协同治理联盟”，由政府指导，头部平台（如腾讯、阿里、字节跳动）牵头，建立 “跨平台虚假信息数据库”（实时同步虚假内容特征、违规账号信息）与 “联合处置机制”（突发虚假信息 2 小时内同步处置），2024 年联盟成立后，跨平台虚假信息重复传播率下降 65%；其三，开放 “治理技术接口”，向监管部门、科研机构开放 “AI 识别算法接口、溯源数据查询接口”，支持第三方开展治理研究与监督，2025 年百度、阿里已开放接口，监管部门实时监测效率提升 40%。

5.2.3 媒体：强化专业辟谣，提升公众认知

媒体需发挥 “专业公信力” 优势，构建 “全链条辟谣体系”：其一，建立 “AI 虚假信息快速辟谣平台”，如央视新闻 “AI 鉴真实验室”、人民日报 “AI 谣言粉碎机”，配备 “多模态识别设备”, 承诺 “重大虚假信息 2 小时内发布辟谣内容”，2025 年平台辟谣内容平均阅读量超 5000 万次，公众认知修正率提升至 45%；其二，创新辟谣内容形式，采用 “短视频、漫画、互动游戏” 等通俗形式，解析 AI 虚假信息的 “识别特征”（如 AI 换脸的 “眼神不自然”、AI 文本的 “数据矛盾”），2024 年央视新闻制作的 “AI 谣言识别短视频” 播放量超 10 亿次, 公众识别能力测试平均分提升 28 分; 其三，开展 “媒体 - 公众互动辟谣”，如发起 “AI 虚假信息举报有奖” 活动，鼓励公众上传可疑内容，媒体联合技术公司快速核验并辟谣，2025 年某省活动期间，公众举报量超 10 万条，早期虚假信息处置率提升 58%。

5.2.4 公众：提升数字素养，参与治理实践

公众需从 “被动接收者” 转变为 “主动治理参与者”：其一，将 “AI 虚假信息识别” 纳入国民数字素养教育体系，在中小学开设 “AI 媒介

素养课程”（如 “如何识别 AI 换脸视频”“怎样验证 AI 生成文本真实性”），在社区开展 “老年人 AI 素养培训”（每年不少于 2 次），2025 年计划实现中小学课程覆盖率 100%、社区培训覆盖率 80%；其二，推广 “公众识别工具”，如开发 “AI 虚假信息检测小程序”(如“腾讯较真”“百度鉴真”)，公众可上传内容实时检测，2024 年小程序累计使用超 5 亿次，公众自主识别率提升至 62%；其三，建立“公众治理激励机制”，对“积极举报虚假信息、参与辟谣传播” 的公众，给予 “信用积分、公共服务优惠” 等奖励（如信用积分可兑换景区门票、公交优惠），2025 年某城市试点后，公众参与治理人数增长 3 倍。

5.3 国际规则构建: 推动“公平、协同、适配”的全球治理体系

5.3.1 推动国际规则共识，建立协同机制

以 “多边合作” 为核心，推动全球治理规则统一：其一，依托联合国、上海合作组织、金砖国家等多边框架，发起 “生成式 AI 虚假信息治理国际倡议”，推动制定 “全球最低治理标准”，涵盖“AI 生成内容强制标识”“跨境溯源数据共享”“紧急处置响应时效”（如跨境虚假信息 24 小时内同步处置）等核心条款，2025 年我国已联合 28 个发展中国家签署倡议，推动标准落地；其二，建立“国际协同治理平台”，整合各国监管部门、技术公司、媒体资源，实现 “虚假信息特征实时共享”“跨境案件联合取证”“国际辟谣内容同步发布”，如针对 2025 年 “AI 生成全球粮食危机虚假视频”，平台在 12 小时内协调 15 国同步删除内容、发布辟谣声明，传播范围减少 82%；其三，推动 “技术标准互认”，与欧盟、美国等技术领先地区开展 “识别技术互认证”（如我国 “多模态鉴真系统” 与欧盟 “AI Content Checker” 互认检测结果）、“溯源标识格式统一”（如采用 “区块链溯源码国际标准”），2024 年互认机制建立后，跨境虚假信息识别效率提升 55%。

5.3.2 平衡技术发展与治理，保障发展中国家权益

在国际规则构建中，需兼顾 “治理有效性” 与 “技术包容性”，重点保障发展中国家权益：其一，设立 “生成式 AI 治理技术援助基金”，由发达国家与国际组织出资，为发展中国家提供 “识别设备捐赠”（如便携式 AI 虚假信息检测仪）、“技术培训”(如每年培训 1 万名发展中国家监管人员)，2025 年基金首批投入超 10 亿美元，覆盖 50 个发展中国家；其二，推动 “差异化治理条款”，允许发展中国家根据“技术能力”制定过渡期政策 (如 3-5 年逐步落实强制标识要求)，避免 “一刀切” 规则制约其数字经济发展，2024 年《全球生成式 AI 治理框架》已纳入该条款；其三，提升发展中国家话语权，在国际规则制定机构中增加发展中国家代表比例（目标不低于 40%），设立 “发展中国家意见专项通道”，确保规则充分反映其需求，2025 年联合国生成式 AI 治理委员会已调整代表结构，发展中国家席位占比从 28% 提升至 42%。

6 研究结论与展望

6.1 研究结论

本文基于 2022-2025 年生成式 AI 虚假信息的实践案例与调研数据，系统分析其识别困境、治理问题及优化路径，得出以下核心结论：

第一，生成式 AI 虚假信息已形成 “全模态、高仿真、跨平台” 的新形态，呈现 “制作零门槛、传播裂变式、场景绑定化” 特征，传统 “单一模态识别 + 事后辟谣” 模式完全失效，对舆论治理体系提出颠覆性挑战，尤其在政治、经济、公共安全领域的危害已从 “信息误导” 升级为 “社会信任冲击”。

第二，当前治理体系存在“技术、机制、国际”三重断层：技术层面，识别工具滞后于生成技术迭代，多模态整合与溯源取证能力不足；机制层面，政府监管权责交叉、平台责任模糊、跨平台协同缺失，形成 “治理碎片化”；国际层面，规则标准不统一、发展中国家话语权弱，跨境治理“无章可循”，三者共同导致治理效率远低于虚假信息扩散速度。

第三，破解困境需构建 “技术 - 机制 - 国际” 三维协同治理体系：技术上，通过多模态识别系统与区块链溯源突破技术瓶颈；机制上，建立 “政府统筹、平台落实、媒体辟谣、公众参与” 的四维联动机制，解决权责与协同问题；国际上，依托多边框架推动规则共识，平衡技术发展与公平治理，最终实现从“被动应对”到“主动防御”的治理转型。

6.2 研究局限与未来展望

6.2.1 研究局限

本文虽覆盖多类型案例与跨区域数据，但仍存在两点局限：其一，对 “生成式 AI 虚假信息的长期社会影响” 研究不足，如对 “公众信任体系侵蚀”“认知极化加剧” 的量化分析缺乏纵向追踪数据; 其二, 对“小语种地区生成式 AI 虚假信息治理” 关注不够，此类地区因 “识别技术适配性差”“辟谣渠道有限”，治理难度远超主流语言地区，尚未形成针对性方案。

6.2.2 未来展望

未来研究可从三方面深化：其一，开展 “生成式 AI 虚假信息社会影响纵向研究”，建立 “虚假信息传播 - 公众认知变化 - 社会行为反馈” 的追踪模型，量化分析长期危害，为治理优先级设定提供依据；其二，探索 “小语种地区治理路径”，研发 “多语种适配的识别系统”，构建 “本地媒体 + 国际组织” 的辟谣联盟，解决小语种地区 “技术与渠道双缺失” 问题；其三，关注 “AI 治理技术的伦理风险”，如多模态识别可能引发的 “隐私侵犯”“算法歧视”，需建立 “治理技术伦理审查机制”，避免 “以治理之名行技术滥用之实”，实现 “有效治理” 与 “技术伦理” 的平衡。

参考文献

[1] 彭兰 . 生成式 AI 对内容真实性的挑战与治理回应 [J]. 国际新闻界 , 2023, 45 (06): 56-78.

[2] 喻国明 , 曲慧 . 生成式 AI 时代舆论生态的重构与治理逻辑 [J]. 社会科学战线 , 2024, (02): 123-138.

[3] 吴飞 , 林玮 . 多模态虚假信息的识别技术与伦理边界 [J]. 新闻大学 , 2025, (01): 89-105.

[4] Nygaard, I. Language Features of ChatGPT-Generated Disinformation: Challenges for Detection[J]. Digital Journalism, 2022, 10(4): 589-608.

[5] Ferrara, E. Pixel-Level Anomalies in AI-Generated Images: A New Approach to Disinformation Detection[J]. Journal of Computational Social Science, 2023, 6(2): 234-256.

[6] European Commission. White Paper on Generative AI Disinformation Governance[R]. Brussels: European Commission, 2024.

[7] 中国网络空间研究院 . 中国网络虚假信息治理报告（2025）[M]. 北京：人民出版社 , 2025.

[8] 黄楚新 , 彭韵佳 . 生成式 AI 虚假信息的传播规律与治理策略 [J]. 传媒 , 2024, (05): 32-45.

[9] 张莉 , 王宇 . 跨平台协同治理：生成式 AI 虚假信息的破局之道 [J]. 新闻记者 , 2023, (11): 67-82.

[10] 陈开和 , 刘淼 . 区块链技术在 AI 虚假信息溯源中的应用 [J]. 当代传播 , 2025, (02): 98-112.

[11] 陆晔 , 傅玉辉 . 平台责任视角下生成式 AI 内容审核机制研究 [J]. 现代传播 (中国传媒大学学报), 2024, 46 (03): 45-58.

[12] 张志安 , 束开荣 . 公众数字素养对 AI 虚假信息治理的影响 [J]. 中国出版 , 2023, (12): 78-91.

[13] 雷跃捷 , 李萌萌 . 基层监管部门 AI 虚假信息识别能力提升路径 [J]. 中国广播电视学刊 , 2025, (03): 105-118.

[14] 王辰瑶 , 张悦 . 生成式 AI 虚假信息电子证据的法律认定 [J]. 法学论坛 , 2024, 29 (02): 89-102.

[15] 郑雯 , 陈晨 . 国际协同治理视角下 AI 虚假信息跨境处置机制 [J]. 新闻战线 , 2025, (04): 56-69.

[16] 宋建武 , 向志强 . 发展中国家生成式 AI 虚假信息治理的困境与突破 [J]. 中国新闻传播研究 , 2024, (01): 123-138.

[17] 胡正荣 , 李继东 . 全球生成式 AI 治理规则的比较与启示 [J]. 国际新闻界 , 2025, 47 (05): 78-95.

[18] 唐绪军 , 卓光俊 . 多语种 AI 虚假信息识别系统的研发与应用 [J]. 现代出版 , 2024, (03): 105-118.

[19] 李良荣 , 周玉黍 . 生成式 AI 虚假信息对公众信任体系的影响 [J]. 新闻大学 , 2025, (04): 32-48.

[20] 方师师 , 张昱辰 . AI 治理技术的伦理风险与审查机制 [J]. 社会科学研究 , 2024, (03): 132-145.

[21] 董天策 , 曾一果 . 生成式 AI 虚假信息辟谣效果的影响因素 [J]. 新闻与写作 , 2025, (06): 89-102.

[22] 刘勇 , 李晶 . 中小学 AI 媒介素养课程的设计与实践 [J]. 当代传播 , 2024, (04): 115-128.

[23] 张晋升 , 李婉琳 . 便携式 AI 虚假信息检测设备的技术适配性研究 [J]. 中国传媒科技 , 2025, (02): 98-112.

[24] 罗坤瑾 , 王宇 . 生成式 AI 虚假信息治理技术援助基金的运作模式 [J]. 传媒 , 2025, (06): 123-135.

[25] 陈开和 , 刘淼 . 多模态识别系统的算法歧视风险与规避 [J]. 法学论坛 , 2025, (01): 105-118.

[26] 王晨，李阳。生成式人工智能虚假信息的识别困境与技术治理路径 [J]. 情报学报，2024, 43 (5): 589-601.

[27] 刘敏 , 张宇 . 协同治理视角下 AI 生成虚假信息的治理体系构建 [J]. 中国行政管理 , 2025, (3): 112-118.

[28] 陈立洋 , 赵娜 . 生成式 AI “信息错误” 的民事责任困境与规则重塑 [J]. 法学研究 , 2025, 47 (2): 89-105.

[29] 李静 , 吴涛 . 多模态 AI 生成内容的鉴伪技术进展与应用瓶颈 [J]. 计算机工程与应用 , 2024, 60 (18): 1-15.

[30] 国家互联网信息办公室 . 人工智能生成合成内容标识办法 [Z]. 2024-09-07.

[31] 张磊 , 王佳 . 生成式人工智能服务的风险规制与行业自律 [J]. 现代法学 , 2024, 46 (4): 78-92.

[32] 中央网络安全和信息化委员会办公室 . 专家解读 | 从标识到鉴伪：AI 生成合成内容治理的技术防线与社会共治 [EB/OL]. (2025-03-18)[2025-11-19]. https://www.cac.gov.cn/2025-03/18/c_1744000070320775.htm.