聚知刊出版社
JZK publishing

浅析 AI 生成虚假信息的法律挑战与协同治理路径

包林强

贵州财经大学 法学院，贵州贵阳，550032；

摘要：生成式 AI 技术突破重塑信息生产范式，其低门槛与隐蔽性催生虚假信息泛滥，威胁公民权益与社会稳定。其治理面临多重困境，即主体责任界定模糊，技术突破加剧监管滞后，法律规制滞后且碎片化。因此需构建多层级协同治理体系，明确技术开发者、平台及用户责任，引入技术风险贡献度评估机制，完善举证责任与归责原则，平衡技术创新与安全风险。

关键词：生成式 AI；虚假信息泛滥；协同治理；责任界定

DOI：10.69979/3029-2700.25.08.064

引言

生成式 AI 正深度重塑信息生产范式，以 DeepSeek、ChatGPT 为代表的模型突破多模态生成局限，仅需少量提示词即可产出逼真内容。其技术低门槛与侵权隐蔽性，催生了虚假信息泛滥的滋生环境。2025 年全国两会，多位代表委员提出 AI 虚假信息传播迅速，尤其是“AI 换脸拟声”和虚假视频等虚假信息的滥用，对公民个人信息和社会稳定构成威胁[1]。虽然我国对人工智能技术的发展持包容审慎态度，但虚假信息治理已突破传统框架，形成涵盖技术研发、传播控制、法律追责的全链路挑战。开发者提供工具、平台搭建渠道、用户实施侵权，各环节耦合放大危害，亟需政府引导多元主体协同治理，通过法律规制、技术防控、伦理规范同步演进，方能在数字时代筑牢信息安全防线，平衡技术创新与社会秩序。

1 生成式 AI 虚假信息治理的多维困境

1.1 责任主体多样化且难以界定责任主体

虚假信息治理的责任体系呈现多主体交织的复杂网络，其责任界定难题源于技术架构与法律规制的错位。我国针对网络虚假信息的监管法规体系虽已初步建立，但仍存在明显的冲突与不足。例如，《AIGC 管理暂行办法》将信息提供者视为网络信息内容的生产者，并要求其履行网络信息安全义务，然而，《网络信息内容生态治理规定》却明确将网络信息内容的生产者定义为制作、复制、发布网络信息内容的组织或个人，并未将技术支持者纳入其中。而《深度合成管理规定》则进一步扩大了责任主体的范围，将深度合成服务提供者和技术支持者均视为责任主体。这种主体界定上的不一致，不仅可能导致责任归属的混乱，还可能因主体责任过严而阻碍生成式人工智能的健康发展[2]。同时，《AIGC 管理暂行办法》还缺乏对用户责任的规定，用户作为生成式人工智能网络生态的重要一环，其生成的内容对于系统运行的反馈和优化具有关键作用，用户责任的缺失无疑会加剧网络虚假信息的风险[3]。

此外，基于“控制理论”确立责任归属虽然提供了基础框架，但技术提供者、服务运营者、内容发布者及用户间的动态交互，导致责任边界呈现流动性特征。技术公司作为生成式人工智能平台的开发者，既掌握模型训练的主导权，又通过 API 接口将生成能力开放给用户，其技术设计缺陷可能引发系统性风险；网络平台作为内容传播的枢纽，在算法推荐机制下对虚假信息的扩散速度具有乘数效应；用户作为生成式人工智能信息的反馈者和信息终端的发布者，其信息的生成主要基于人类反馈的强化学习，信息发布时其主观恶意与认知局限形成责任判定的伦理困境[4]。

1.2 生成式人工智能技术突破下的治理矛盾与协同防控机制构建需求

以 DeepSeek 平台为例，其自研的 DeepSeek-R1 模型采用组相对策略优化算法与混合专家架构，通过三阶段训练“冷启动-强化学习-全场景优化”实现了长链推理与多语言支持的技术突破。该模型融合奖励机制、拒绝采样、自进化学习及知识蒸馏等创新技术，在提升推理效率与部署灵活性的同时，也暴露出深层治理矛盾，算法决策的“黑箱”特性导致监管难以穿透技术屏障，模型训练中的数据偏差可能衍生歧视性输出，强化学习机制下的自我进化存在伦理失控风险，而长链推理过程中的不可控性更可能生成逻辑自洽的虚假信息，对社会认知体系构成系统性威胁[5]。

二元治理结构在 AI 虚假信息治理中呈现理论上的协同优势，但实践层面存在显著实施矛盾[6]。政府规制通过立法确立责任边界，欧盟《人工智能法案》尝试对高风险 AI 系统实施严格监管，我国《生成式 AI 服务管

理暂行办法》探索算法备案制度。市场机制则依赖企业技术自律，例如 DeepSeek 等平台通过内容安全过滤、人工审核介入等机制进行风险缓释。然而，算法技术的复杂性与法律规制的滞后性形成制度性错位，例如现行法律框架难以准确界定生成式 AI 的“产品”属性，算法决策过程缺乏可解释性导致责任认定困难，企业技术治理的自愿性特征削弱监管效能，而跨平台虚假信息的传播链条更放大了治理碎片化问题。这种技术理性与法律秩序的张力，要求构建政府监管、企业责任与技术治理的协同机制，通过算法审计、算法监督、伦理审查、多方共治等制度创新，实现虚假信息风险的体系化防控。

1.3 生成式 AI 虚假信息法律规制的双重困境

生成式人工智能引发的虚假信息风险正对传统法律规制体系形成显著冲击，其损害已突破传统法益保护框架，呈现出复合型治理难题。在私法领域，虚假信息通过深度伪造技术侵害人格权益的路径愈发隐蔽，常引发名誉权、肖像权、隐私权的多重叠加侵害。民事法律层面，传统侵权责任体系面临技术挑战：算法自主决策场景下，过错责任原则难以适用，受害者难以证明 AI 开发者或使用者的主观过错；人格权保护框架存在滞后性，如《民法典》第 1023 条将声音权益保护参照肖像权的权宜安排，未能有效应对 AI 拟声技术带来的新型侵权风险[7]，生物特征信息的独立保护需求亟待立法回应；电子证据规则与生成式 AI 的技术特性存在制度错位，导致受害者面临举证困难，尤其在算法黑箱场景下，虚假信息源于模型训练数据偏差或算法缺陷时，责任主体认定更加复杂。

行政监管体系则呈现碎片化特征，规范效力位阶冲突凸显部门规章与上位法的衔接困境。《生成式人工智能服务管理暂行办法》虽引入产品质量责任条款，但 AI 生成内容是否属于产品，是否属于《产品质量法》规制范畴存在理论争议，其第 9 条在司法适用中可能陷入解释困境；《深度合成管理规定》与《算法推荐管理规定》在生成式 AI 治理上存在制度竞合，前者聚焦内容深度伪造治理，后者强调算法透明度监管，二者在责任主体认定、归责原则适用等核心要件上缺乏规范协同。多部门监管权责配置模糊导致治理真空，网信、工信、公安等部门在生成式 AI 全生命周期监管中的职能边界尚未厘清，面对“AI 换脸拟声”等复合型虚假信息侵害，现有规范呈现碎片化应对特征，过度依赖基础民法条款的零散保护，难以形成有效规制合力。

2 AI 虚假信息风险治理治理的法治路径

2.1 厘清虚假信息治理的主体责任

生成式 AI 技术架构与法律规制的错位矛盾，导致虚假信息治理责任体系面临深层挑战。传统控制理论虽已有责任划分基础，但 AI 多主体交互特性使责任边界动态流动，技术开发者作为算法架构核心设计者，主导模型训练又开放 API 接口，其技术缺陷可能引发系统性风险；服务运营者通过算法推荐机制对虚假信息扩散产生乘数效应；内容发布者与使用者作为传播起点，其主观认知直接影响责任认定，普通用户可适用过错推定原则。这种多主体责任链条要求分层治理机制，技术开发者需承担算法伦理设计责任，通过技术备案与伦理评估确保模型鲁棒性；平台需建立内容溯源系统，对未标注 AI 生成内容或篡改标识行为实施梯度追责。需注意的是，AI 强化学习特性使模型与用户协同进化，当虚假信息自我增强时，技术架构成为滋生环境，开发者与平台应当共同承担责任。

法律规制创新需突破传统模式，引入“技术风险贡献度”评估机制，构建包含技术可控性、审核有效性、用户引导性等维度的量化指标体系，明确各主体责任权重。同时建立多主体协同治理框架，技术公司提供算法审计支持，平台开放审核数据池，科研机构构建虚假信息特征库，形成覆盖“生产-传播-监管”全链路的技术联防体系。这种责任界定机制创新，既能化解责任模糊困境，更可推动 AI 技术向可控、可信、可用方向演进，通过量化评估与协同治理实现技术风险与法律规制的动态平衡。

2.2 构建虚假信息全链条治理体系

构建全链条形式的分层级虚假信息治理体系[8]，需构建“政府监管-平台自治-开发者自律-用户参与”的协同治理网络[9]。政府监管层面，应构建“准入审查+动态评估”框架，建立生成式 AI 服务开发者白名单制度，审查模型架构、训练数据、算法逻辑合理性及合法性，重点关注数据隐私、算法偏见、伦理风险。对服务平台实施分级管理，对未建立内容审核体系或审核失效的平台采取强制措施，并建立开发者责任追溯机制，要求技术文档备案含算法伦理影响评估报告，对高风险模型实施“沙箱测试”准入。平台自治方面，应构建“AI 过滤+人工复核+用户协同”立体审核体系，大型社交媒体部署多模态内容风控矩阵，建立热门话题 AI 巡查机制，强化身份验证机制，推行“真人账户”优先推荐制度，建立账户信用分级体系。开发者自律层面，构建“技术对抗+人工介入+备案溯源”防控链条，建立 AI 反制 AI 技术体系，部署对抗生成网络检测虚假信息，设置双重审核机制，建立内容生成溯源机制，实现全链路可追溯。此外，还需构建“数据合规+模型监测+伦理校准”

动态治理机制，建立训练数据全生命周期监管体系，健全模型反馈机制，设立跨学科伦理委员会，形成“监测预警-分级处置-迭代优化”自主进化系统。这种多层级事前防控机制，通过技术管控与制度约束深度融合，既保障技术创新活力，又筑牢伦理法律底线，共同构建适应人工智能发展的现代治理体系。

2.3 构建虚 AI 假信息治理法律框架的多元平衡路径

构建完整的虚假信息治理法律框架，应以精细化责任分配与技术创新激励平衡为核心。在责任主体界定上，需深化《生成式人工智能服务暂行办法》中“生产者责任”内涵，明确技术开发者、服务运营者、内容发布者的三级责任体系。开发者承担算法伦理设计责任，通过技术文档备案、伦理影响评估等机制确保模型稳健性；运营者建立内容溯源追踪系统，对未标注 AI 生成内容或恶意篡改标识行为实施梯度追责；发布者主观认知影响责任认定，普通用户适用过错推定原则，恶意传播者须承担严格责任[10]。

归责原则设计需体现技术特性与法律原则的融合创新。对于因技术缺陷引发的错误，应适用无过错责任原则，类比产品责任的处理方式，要求技术提供者承担强制性的质量担保责任；对于数据错误或用户诱导导致的错误，则适用过错推定责任，要求责任主体自证无过错，以在技术创新与责任承担之间建立动态平衡。这种分层归责机制既强化技术提供者的质量把控责任，又避免对技术创新产生过度抑制。因果关系认定机制需引入程序法创新。借鉴欧盟《人工智能法案》的举证责任倒置规则，建立服务提供者、开发者的连带责任机制。受害人仅需证明“信息错误”与损害之间存在合理关联，服务提供者则需承担反证无因果关系的责任。通过标识义务、风险评估和备案机制，降低利益损害者的举证难度，使责任界定更具操作性。

此外，协议类型法定化需规范市场秩序，技术提供者和服务提供者签订标准化注册协议和服务合同范本，明确质量标准、错误处理流程、数据使用边界等条款，为消费者提供权益保障依据，为监管部门提供监督抓手。在事后追中损害计算规则的完善需突破传统侵权法局限，建立量化损害赔偿标准，引入动态评估机制，确保赔偿数额与损害程度匹配。建立惩罚性赔偿制度，对恶意传播虚假信息行为形成有效威慑。但对于间接损害，不应无限扩大，应当采取“合理预期收益+法院自由裁量确定”的方法进行认定[11]。

3 结语

在生成式 AI 重塑信息生态的当下，AI 虚假信息治理已演变为涵盖技术、法律、伦理的全域挑战。面对深度伪造技术引发的认知污染，需构建政府、平台、开发者、用户协同共治的治理网络，通过责任主体精细化划分、技术风险贡献度评估、举证责任倒置等制度创新，实现法律规制与技术防控的双向赋能。治理体系升级应贯穿模型训练、内容生成、传播扩散全周期，融合事前审查、事中监测、事后追责的动态防控机制，在保障技术创新的同时筑牢伦理底线，推动 AI 技术深度赋能社会各领域，最终达成技术理性与社会价值的和谐统一。

参考文献

[1]文丽娟.代表委员建言加强AI虚假信息治理[N].法治日报,2025-03-06(007).
[2]张素华,李凯.生成式人工智能虚假信息风险与治理研究[J].学术探索,2024,(07):129-140.
[3]刘心怡.AI 生成虚假信息的法律规制[J].服务外包,2025,(03):42-44.
[4]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,(05):34-54.
[5]邓建鹏,赵治松.DeepSeek 的破局与变局:论生成式人工智能的监管方向[J/OL].新疆师范大学学报,2025,(04):1-10.
[6]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(04):142-156+160.
[7]朱溯蓉.语音生成式人工智能的风险与治理[J/OL].电子科技大学学报,1-10.
[8]苏宇.大型语言模型的法律风险与治理路径[J].法律科学,2024,42(01):76-88.
[9]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(03):112-120.
[10]黎明,张益欣.数字时代的谎言：虚假信息的识别与法律治理的向度[J].昆明理工大学学报,2025,25(01):1-9.
[11]郭金良.生成式人工智能服务中“信息错误”的民事责任[J/OL].政法论坛,2025,(02):25-35.

作者简介：包林强，硕士研究生，方向为人工智能法学