AIGC 时代网络舆情风险的政府治理

李怡
天津商业大学，中国·天津 300134

摘要：人工智能生成内容（Artificial Intelligence Generated Content，AIGC）技术的快速发展，正深刻改变网络舆情风险的治理逻辑，并对政府治理体系形成新的压力。本文以“技术－制度－价值”为分析框架，梳理了2017年至2026年间发布的20份国家战略文件。研究发现，现行治理体系虽已在技术防控、制度规制和价值规范三个维度有所布局，但各维度在实践中均面临现实挑战，制约了治理效能的充分释放。基于此，本文提出敏捷防御、弹性规制、嵌入引领的治理取向，以期提升AIGC时代网络舆情风险治理的实效。

关键词：人工智能生成内容；网络舆情风险；政府治理；技术－制度－价值

Government Governance of Online Public Opinion Risks in the Era of AIGC

Li Yi
Tianjin University of Commerce, China Tianjin 300134

Abstract: The rapid development of Artificial Intelligence Generated Content (AIGC) technology is profoundly transforming the logic of managing online public opinion risks and placing new pressures on government governance systems. Using a "Technology-Institution-Value" analytical framework, this paper reviews 20 national strategic documents issued between 2017 and 2026. The study finds that while the current governance system has already established measures across the three dimensions of technological prevention, institutional regulation, and value-based norms, each dimension faces practical challenges in implementation, limiting the full realization of governance effectiveness. Based on this, the paper proposes a governance approach characterized by agile defense, resilient regulation, and embedded leadership, with the aim of enhancing the effectiveness of online public opinion risk governance in the AIGC era.

Keywords: Artificial intelligence generated content; Online public opinion risks; Government governance; Technology-institution-value

0 引言

人工智能生成内容（Artificial Intelligence Generated Content，AIGC）技术的快速普及，正在深刻改变网络舆情的生成逻辑与传播生态。深度伪造、跨模态生成、个性化内容分发等应用日趋日常化，而现有网络舆情治理体系仍以线性规则和属地管理为基本框架，在应对技术快速迭代和跨境风险扩散时显得力不从心。从既有研究来看，学者们大多聚焦于技术检测、制度规制或价值引导中的某一维度，对三者之间的互动关系和整合路径关注不足。基于这一问题意识，本文尝试构建“技术－制度－价值”(Technology-Institution-Value，TIV）三维整合分析框架，对2017-2026年间中国政府治理AIGC网络舆情风险的相关实践进行历时性梳理，试图回答三个核心问题：当前治理体系在实践中呈现出怎样的整体面貌？现有做法面临哪些现实困境？又该如何建立一套更能回应技术特性的治理路径？

1 分析框架与研究方法

1.1 分析框架

传统的公共治理研究一直存在着技术决定论与制度中心主义的二元对立，前者把技术当成提升治理效率的一种外生工具，后者则强调规则与组织的刚性约束。AIGC时代，网络舆情风险的治理面临技术迭代与制度滞后的结构性矛盾，只靠单一维度已无法揭示其治理全貌。基于此，本文整合技术治理理论[1]、制度主义理论[2]与公共价值管理理论[3]，借鉴胡登峰等的“技术－制度－价值”三维框架[4]，构建AIGC时代网络舆情风险政府治理的分析框架。其中，技术维度属于工具理性范畴，是舆情治理的硬件基础。制度维度属于规则约束的范畴，包含了法律法规与监管权责等正式规则。价值维度属于目标导向的范畴，锚定了网络意识形态安全、算法公平等核心价值。这三个维度互构共生、协同支撑，共同构成AIGC舆情治理的整合性分析。

1.2 研究方法

本研究采用政策文本分析法，围绕技术、制度、价值三个方面，对 2017 年至 2026 年间中国政府治理 AIGC 网络舆情风险的相关政策进行梳理与归类分析。政策样本的挑选遵循以下原则。一是内容相关性。即所选政策文本须直接或间接规制 AIGC 作为风险源引发的网络舆情风险，剔除仅聚焦技术研发、产业标准体系建设等关联度较低的文件。二是效力权威性。本研究只纳入国家层面权威机构发布的规范性文件。三是时间跨度。本研究将 2017 年至 2026 年作为考察时段，以 2017 年发布的《新一代人工智能发展规划》为 AIGC 治理的政策起点，试图覆盖技术防御、立法建制到价值引领的政府治理实践过程。经过筛选，最终选取了 20 份政策样本，所有文本均取自中央人民政府官网与国家互联网信息办公室官方平台，完整的政策清单见表 1。

表 1 政策样本清单

序号	政策名称	发布时间	序号	政策名称	发布时间
P001	《新一代人工智能发展规划》	2017.7.8	P011	《网络暴力信息治理规定》	2024.6.12
P002	《新一代人工智能治理原则——发展负责任的人工智能》	2019.6.17	P012	《人工智能安全治理框架》1.0 版	2024.9.9
P003	《网络音视频信息服务管理规定》	2019.11.18	P013	《网络数据安全管理条例》	2024.9.24
P004	《网络信息内容生态治理规定》	2019.12.15	P014	《互联网信息服务管理办法》（2024年修订）	2024.12.6
P005	《关于加强网络文明建设的意见》	2021.9.14	P015	《网络安全技术 人工智能生成合成内容标识方法》	2025.2.28
P006	《新一代人工智能伦理规范》	2021.9.25	P016	《人工智能生成合成内容标识办法》	2025.3.7
P007	《互联网信息服务算法推荐管理规定》	2021.12.31	P017	《人工智能安全治理框架》2.0 版	2025.9.15
P008	《互联网用户账号信息管理规定》	2022.6.27	P018	《中华人民共和国网络安全法》（修正）	2026.1.1（施行）
P009	《互联网信息服务深度合成管理规定》	2022.11.25	P019	《人工智能科技伦理审查与服务办法（试行）》	2026.3.20
P010	《生成式人工智能服务管理暂行办法》	2023.7.10	P020	《人工智能拟人化互动服务管理暂行办法》	2026.4.10

作者整理

2 政府治理的实践图谱与现实困境

2.1 实践图谱

2.1.1 技术维度

技术维度已经形成了从源头评估到过程标识再到闭环迭代的全流程防控。按照政策要求，治理责任需要追溯到算法开发者，因此模型从研发阶段就要接受安全评估。P010 规定训练数据源必须合法[5]，P009 则明确，涉及生物识别信息或者国家安全、社会公共利益的模型工具，必须依法开展安全评估，这是对高风险技术应用设置了准入门槛。与此同时，为了建立可追溯的数字身份体系，政策要求对 AIGC 内容进行标识并留存日志，治理思路从单纯的过滤虚假信息转向规范化管理。另外，政策还要求服务提供者把处置过程中的漏洞反馈给模型训练环节，进行对抗加固，形成一种能够自适应的风险防御机制。

2.1.2 制度维度

制度层面已经搭建起层级规则、梯度责任和协同监管的多维框架。现行体系大致分三层：P001 定下了治理的基本方向，P009、P010 等部门规章，把原则变成了具体的合规要求，P016 这类国家标准，则将监管要求统一成了技术参数。同时，政策围绕基础模型开发、微调、API 调用、终端传播这些环节，设计了分层追责的链条，并用穿透式监管和多层次处罚来压实合规责任。另外，治理架构上由网信部门牵头，多个部门共同参与，形成了联合治理的格局。

2.1.3 价值维度

价值层面上，底线约束、伦理引领与多元共治的规范架构已初步成型。政策以正面清单和负面清单共同约束价值行为。P010 将弘扬社会主义核心价值观纳入立法目标，为价值判断提供了基础准则，负面清单主要是明确严禁利用 AIGC 技术制造虚假新闻或操纵公众舆论。同时，政策通过软性价值引领推动科技向善具象化与治理主体多元化，P006 将公平公正、保护隐私等原则细化为算法偏见防范、数据授权等具体要求，并鼓励行业组织、企业、科研机构和公众共同参与治理。

综合以上维度，我国 AIGC 舆论治理的框架：技术工具、制度规则和公共价值，大体上已经搭建起来。但政策进入执行环节后，技术、制度和价值这三个方面都暴露出了弱点。更为重要的是，一个方面的缺陷在实践中会被其他方面的不足放大，让治理变得更为棘手。下文将围绕这三个方面，分析政府面临的实际挑战。

2.2 现实困境

2.2.1 技术维度：检测溯源能力与风险演化速度的错配

多模态检测存在明显盲区，伪造技术迭代速度远超检测特征数据库的更新能力。2025 年厦门化工厂爆炸谣言事件中，造谣者借助 AIGC 工具批量制造煽动性短视频，平台内容审核未能及时拦截，导致流言四起。在这种情况下，若相关部门缺乏实时多模态分析能力，便会错失干预时机。标识追踪技术同样面临困境，黑灰产可轻易绕过表层显式标识，通过裁切或覆盖就可以将其清除，溯源链条随之断裂。

2.2.2 制度维度：权责边界模糊与协同监管的困境

制度层面首先面临的是责任不清的问题。在复杂场景中，各方的责任边界仍然很模糊。若用户微调开源基础模型后生成有害信息，开源社区、微调主体、托管平台与终端用户各方的责任边界没有进行清晰的划分，追责自然存在困难。另一问题是部门协作受阻，监管资源分配不合理。当 AIGC 舆情风险与金融、政务等行业深度绑定时，各部门在职责交叉的地方研判标准不统一、信息共享不及时，公安涉案音视频数据、工信备案日志与网信舆情样本因分类标准不同和部门壁垒无法融合训练。特别是跨境信息流动也为属地管辖带来难题。

2.2.3 价值维度：伦理规范悬浮与共识凝聚的双重挑战

落到价值层面，AIGC 对算法茧房的放大效应值得警惕。其内容输出迎合用户偏好，情绪化与偏见性明显，主流叙事空间被挤压；批量伪造评论、篡改民调等行为则制造虚假共识，公共讨论从事实辩论滑向算法操纵，社会撕裂风险加剧[6]。科技向善原则缺少硬性约束，执行层面流于形式。平台追逐流量变现，算法以点击率为优先，信息多样性被边缘化。2025 年“清朗”专项行动报告指出，不少平台对 AI 生成内容博流量的行为持默许态度，即便某些内容存在争议或低俗倾向，道德准则仅停留在纸面。

3 政府治理的优化路径

3.1 技术维度：敏捷防御与深度溯源

构建国家级红蓝对抗攻防演练平台，统筹网信、工信等部门算力资源，组建常态化风险模拟平台，红队利用最新开源模型模拟生成高仿真舆情攻击内容[7]，蓝队同步迭代多模态识别模型与风险特征库，建立全国统一威胁情报共享渠道，压缩检测技术迭代滞后周期。同时研发具备鲁棒性的跨媒介水印与区块链全程存证系统，开发可抵御裁切、转码、翻拍等攻击的隐形水印，在 AIGC 模型输出端口接入区块链存证模块，将内容生成的模型指纹、时间戳、初始用户等关键信息上链，形成不可篡改的内容 DNA[8]，为穿透式追责提供技术支撑。

3.2 制度维度：弹性规制与穿透式监管

将监管范围扩展至嵌入 AIGC 生成功能的高风险产品，厘清各方权责边界，破解技术链条中责任被稀释的难题。推行风险分级管理与监管沙盒机制，对用于新闻、金融、教育等高风险的 AIGC 服务实施严格前置安全评估和持续审计，对娱乐、办公辅助等低风险场景采取备案制或事后监测为主的方式，在可控环境中允许创新应用试错，平衡技术创新与安全底线。建设国家级 AIGC 风险数据共享平台[9]，推动网信、公安、工信等部门间涉 AIGC 风险数据的标准化共享与融合分析，技术系统将识别出的风险线索实时同步至平台实现精准分拨，同时依托多边国际组织推动 AIGC 治理标准共建，建立境外有害 AIGC 网络舆情线索的移送与证据协查合作机制。

3.3 价值维度：机制嵌入与价值引领

分层分类推进全民算法素养教育，将 AIGC 相关内容纳入中小学通识与高校公共课程，提升公众辨别深度伪造和规避算法茧房的能力。同时设立国家级人工智能研发项目，探索算法技术在意识形态安全建设中的应用路径，建立“AI+ 政策、教育、文化、法律”等运行模式，巩固我国在意识形态领域的工作领导权、管理权与话语权[10]。构建伦理合规与市场激励相衔接的硬性约束机制，将算法公平性、反极化设计原则和伦理审查体系纳入 AIGC 安全评估的强制指标，把平台治理成效与伦理合规记录纳入市场准入与融资考核体系，引导企业将科技向善理念嵌入产品研发全流程。

上述三个维度并非彼此孤立，而是双向关联，并产生协同效应。技术维度的敏捷防御既为制度性监管提供了必要工具，同时也为价值层面相关举措提供了可验证的技术支持。制度维度的灵活监管在界定技术应用的法律边界的同时，将伦理要求转化为具有约束力的合规义务，从而成为连接价值与技术的不可或缺的纽带。价值维度的嵌入引领通过伦理审查来确定技术的发展方向，又通过广泛达成共识来降低制度执行的阻力。由此，技术赋能、制度转化与价值校准形成正向循环，打破“技术滞后 - 制度缺位 - 价值悬浮”的困局，实现风险前置防控、跨主体协同与价值共识培育的多重目标。

4 结语

本文基于“技术 - 制度 - 价值”三个维度，将 AIGC 舆情治理整合到同一分析框架下，揭示了治理体系政策文

本完备与执行效能不足之间的结构性落差，为理解AIGC时代政府治理的复杂性提供了整合性视角，以期实现AIGC时代网络舆情风险政府治理效能的提升。

参考文献：

[1] 刘永谋，李佩. 科学技术与社会治理：技术治理运动的兴衰与反思[J]. 科学与社会，2017(02):58-69.

[2] [美]道格拉斯·C. 诺斯:《制度、制度变迁与经济绩效》，刘守英译，上海：生活·读书·新知三联书店上海分店，1994:5-6.

[3] 昌诚，张毅，王启飞. 面向公共价值创造的算法治理与算法规制[J]. 中国行政管理，2022(10):12-20.

[4] 胡登峰，吴昊，王佳怡. 人工智能时代政府适应性治理研究——基于技术—制度—价值三维分析框架[J]. 学术月刊，2025,57(06):83-94.

[5] 张凌寒. 筑牢深入推进“人工智能+”行动的政策制度之基[J]. 中国经贸导刊，2026(5):13-15.

[6] 梁循，王涵玉，张森森等. 人工智能生成内容对舆情生态的影响、挑战和治理[J]. 管理世界，2025,41(6):129-158.

[7] 何雅妍，周保民. 人工智能“对抗式治理”的创新转向——基于红队测试技术政策的多国别考察[J/OL]. 科学学研究，2026:1-20.

[8] 肖萍，鄢琦昊. AIGC时代网络舆情风险的敏捷治理：适配逻辑、现实困境与实现路径[J]. 求实，2026(2):68-78+107.

[9] 谢耘耕，张伶聪. AIGC驱动下舆论学研究方法的范式重塑与路径探索[J/OL]. 新媒体与社会，2026:1-15.

[10] 黄楚新，贺文文. AIGC时代网络舆情的新特征与治理新对策[J]. 新闻论坛，2024,38(6):7-9.

作者简介：李怡（2000.05-），女，汉族，河南洛阳人，硕士研究生，研究方向：公共危机与应急管理。