法律·科技·社会

LAW TECHNOLOGY SOCIETY

第 5 卷 第 3 期
总第 27 期
2025. 05

香港星源出版社
STAR SOURCE PUBLISHING

3058-1742

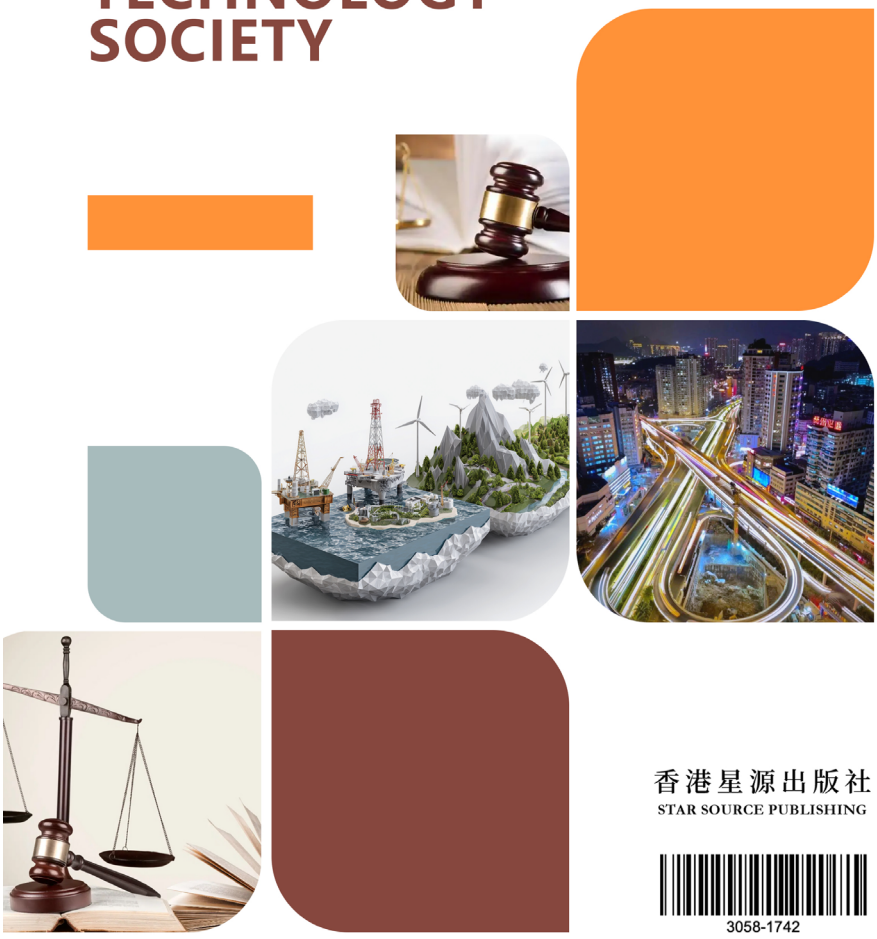
STAR SOURCE PUBLISHING
香港星源出版社

法律•科技•社会　　Law • Technology • Society

2025 年 05 月　　第 27 期

法律•科技•社会

LAW
TECHNOLOGY
SOCIETY



香港星源出版社
STAR SOURCE PUBLISHING

3058-1742

出版社信息

主管：香港星源出版社

主办单位：香港星源出版社

主编：王耀东

执行主编：周翊凡

社内编辑：

陈子昂　赵奕辰　温子墨　王奕彤
宋　靖　黄思源　江语珊　欧阳琳
贺嘉玮　黎芷蕊　李　熙　蒋浩然
曾雅琴　罗　银　龚昱哲　程梓恒
胡　桐

网址：https://hksspub.com/

电话：+852 6855 8145

邮箱：hksspub2022@163.com

刊期：双月刊

目 录 CONTENTS

STAR SOURCE PUBLISHING
香港星源出版社

人工智能操控舆论的风险分析与治理对策研究

林　泽　宇
（北京　北京邮电大学　100876）

摘要：

随着生成式人工智能（AIGC）技术的发展，舆论操控方式呈现出智能化、自动化的新趋势。借助自动文本生成、深度伪造、推荐算法等手段，人工智能可被用于大规模制造虚假信息、操纵情绪走向、影响公共意见结构，带来较大的信息安全风险与治理难题。本文系统梳理了 AI 操控舆论的主要技术路径，包括内容生成、虚假账号集群、水军自动化运行与推荐算法的引导效应，剖析其在传播速度、情绪感染力与识别难度方面的风险特征。

通过对典型案例的分析，揭示当前操控行为在商业营销、政治宣传与危机公关中的应用态势。进一步指出了当前治理体系中存在的技术识别滞后、法律责任不明、平台监管机制缺位与国际协同困难等问题。在此基础上，提出构建多元协同治理体系的策略建议，包括完善内容溯源机制、细化法律责任边界、提升平台自律透明度与公众媒介素养，并探讨了跨国协作的可行路径。研究旨在为人工智能语境下的网络舆论安全治理提供理论基础与实践启示。

关键词：人工智能、舆论操控、信息安全治理、算法推荐机制、虚假信息传播

一、引言：当 AI 成为“舆论放大器”

1.1 新技术背景下的网络舆论变化

随着人工智能技术特别是生成式人工智能（AIGC）的快速发展，网络舆论的生成、扩散与引导方式发生深刻变革。AI 工具能够在极短时间内自动生成文本、图像和视频，显著提升了信息传播效率，但同时也为舆论操控提供了更为精准和隐蔽的手段。在大数据、推荐算法和社交媒体平台联合作用下，网络舆情的传播呈现出加速化、情绪化与极化趋势。平台算法“信息茧房”效应与“同温层”结构加剧了公众意见的群体分化，引发社会对“操控式传播”的深层关注。

1.2 舆论操控方式的智能化转型

相较于传统的人工水军和舆论炒作行为，AI 驱动下的舆论操控具备更强的规模化与技术隐蔽性。例如，通过深度伪造生成虚假视频、自动化生成成千上万条文本评论、智能识别用户画像精准引导信息接受行为，已成为部分政治组织、商业机构或网络团体实施舆论干预的重要方式。这类操控不仅加剧信息失真，还可能引发公共恐慌、损害社会信任结构，构成对网络空间治理的严峻挑战。

1.3 国内外研究现状与存在问题

当前国内外学者对 AI 在舆论传播中的角色已有一定研究，但现有研究普遍存在以下不足：一是对生成式人工智能操控舆论的具体手段与案例分析不足；二是对风险特征与影响机制的动态演化缺乏系统梳理；三是治理对策中多停留于理论构想，缺乏操作层面上的应对策略，尤其在平台责任机制和跨国治理协作方面的探讨仍显不足。

1.4 研究目的、方法与结构安排

本研究旨在系统分析人工智能操控舆论的风险类型与传播机制，探讨其在舆论场中的应用方式与治理挑战，进一步提出切实可行的治理路径。文章采用文献分析法与案例研究法相结合的方法，选取典型平台与事件作为分析对象，结合舆论传播理论与技术特征进行交叉研判。文章结构如下：第二章梳理人工智能介入舆论传播的技术路径；第三章分析智能操控舆论的风险类型与识别难点；第四章剖析典型案例中的操控策略与扩散机制；第五章提出舆论治理的路径建议与政策启示；最后总结研究发现并提出未来研究方向。

二、人工智能如何影响舆论走向？——操控机制解析

2.1 生成式 AI 技术与内容自动生成能力

生成式人工智能（Generative AI）是近年来迅速发展的技术分支，具备自动生成文本、图像、音频与视频内容的能力。基于大型语言模型（如 GPT）、文本生成系统、图像合成模型（如 Stable Diffusion）等技术，AI 可在短时间内输出大量逻辑连贯、情绪导向性强的内容。

这一能力为信息传播带来极大便利，但也降低了

信息操控的技术门槛。借助生成式 AI，舆论操控者无需复杂编程技能，即可批量制作评论、文章、虚假新闻或深度伪造内容，从而在社交平台上制造“议题假象”或引导群体情绪 [5]。

2.2 常见操控方式及实现路径

AI 参与舆论操控的路径主要包括三种方式：一是自动评论与水军模拟，利用 AI 生成自然语言文本批量发帖，模拟真实用户言论，影响讨论方向；二是虚假身份生成与账号操控，通过 AI 合成头像、构造人设，形成“假账号矩阵”实现集群操作；三是多模态深度伪造，通过合成视频、语音与图像干扰受众判断。这些方式往往伴随自动发布与传播脚本，实现内容精准投放，提升舆论干预的隐蔽性与操控力 [2]。

2.3 信息茧房与推荐算法的引导效应

算法推荐系统作为现代社交平台的信息分发核心，其基于用户兴趣和行为轨迹进行内容推送。虽然提高了用户体验，但也可能导致“信息茧房”效应——用户只接触与自身立场一致的信息，从而加剧观点极化。这一机制在 AI 舆论操控中被广泛利用：操控者通过测试平台算法偏好，主动设计内容以触发推荐机制，最终实现特定舆论的放大或压制。例如热点话题操控中，相关内容通过点赞、转发和评论策略被快速推向信息流首页，形成“议题绑架”效应 [1]。

2.4 舆论操控中的实施主体与组织方式

人工智能操控舆论的背后，往往存在多样化的实施主体与分工协作机制。主要主体包括：政治性机构（如境外势力、极端组织）、商业操盘团队（通过数据营销与竞品抹黑操控舆情）以及部分网络舆情公司。操控过程通常包括议题设计、内容生成、账号部署与传播引导四个阶段，有时还与算法优化团队或数据公司形成灰色合作链条。这种跨组织的协作机制使得舆论干预操作更具专业性与持续性，增强了治理难度 [4]。

三、操控风险表现及典型案例分析

3.1 虚假信息制造与内容变形传播

人工智能特别是 AIGC 具备高效的内容合成能力，能够快速生成具“真实感”的文本、图像与视频，使得信息造假更具迷惑性。在舆论操控中，这类 AI 内容被广泛用于制造“似是而非”的新闻、评论与公告，模糊信息真伪边界。

更复杂的是，虚假信息在传播过程中会经历语义扭曲、主观解读和情绪放大等再加工路径。尤其在多平台、多轮次转载后，原始语境易被遗失，虚假内容被赋予新的意义，形成传播“变形链”，扩大误导性影响 [5]。

3.2 情绪放大与用户认知误导

情绪驱动是 AI 舆论操控的关键策略。通过算法识别用户心理特征，系统可定向推送具煽动性的话语，持续刺激用户情绪反应。愤怒、恐惧、焦虑等负面情绪被优先放大，从而实现舆论引爆。

这种机制不仅影响个体判断力，还重构集体认知。例如，在敏感事件中，AI 生成的大量“带节奏”评论会引发网民跟风式愤怒，促使群体陷入非理性舆论漩涡，偏离事实核查与公共理性 [3]。

3.3 商业、公关与政治领域的操控目的

AI 舆论操控已广泛渗透至多类应用场景。在商业营销中，企业通过“刷量”、“伪评”来左右用户感知，操控商品口碑。在危机公关中，AI 账号矩阵用于“洗白”品牌或压制负面话题，形成“议程压制”效果。

更具挑战性的是政治操控行为。部分组织通过跨国平台操作 AI 舆情账号，发布深度伪造的政治言论，意图影响公众选举选择或制造社会撕裂。这种操控往往隐藏在“群众声音”的外衣下，难以快速识别。

3.4 案例一：国内舆情事件中的 AI 操控现象

在某高校“研究生退学风波”中，原始贴文曝光后，一小时内多个平台同步出现高密度评论，内容语气趋同，措辞结构重复。进一步分析发现，多数评论为 AI 自动生成账号发布，用户头像为虚构人像，IP 集中。

该事件的操控路径包括：话题发酵前期利用 AI 发帖炒热，情绪上升阶段生成极端言论激化对立，最后通过多平台扩散制造“公众一致愤怒”假象。尽管事件源点合理，但被操控后演变为广域网络攻击，偏离理性讨论轨道 [4]。

3.5 案例二：海外社交平台中的舆论干预行为

2020 年美国大选期间，Twitter 曾通报约 7,000 个与某国家相关的“自动化政治账户”被清理，这些账户大量使用 AI 生成文本与图像发布假新闻、候选人攻击材料及选民煽动内容，意图扰乱选民信息判断。

AI 操控的关键机制包括：一是通过“账号矩阵”构建表面多样的舆论氛围，二是借助平台推荐算法迅速扩大信息影响力，三是通过图像 + 文本“混合输入”提升内容可信度。最终导致部分议题在舆论场呈现出高度极化甚至反转倾向 [5]。

四、治理困境与技术伦理挑战

4.1 舆论操控识别的技术盲区

尽管 AI 内容生成带来传播便利，但也极大削弱了平台与监管部门对信息真伪的识别能力。传统的舆情监测系统多依赖关键词匹配、账号行为识别等技术

手段，但面对不断演化的生成式模型，如 ChatGPT、Midjourney、Sora 等，其所产出的内容语义连贯、形式多样，已难以用既有规则甄别。

AI“模仿真实用户”的能力日益增强，其语言风格、时间分布、互动行为等已接近自然用户群体，导致“异常检测”与“假账号识别”难度大幅提升。技术治理一旦滞后，将使操控行为潜藏更深，甚至造成监管空窗期 [1]。

4.2 平台治理责任落实难

社交媒体平台作为舆论扩散的核心载体，在信息推荐、内容审核与风险预警中承担关键角色。然而当前平台治理中普遍存在以下问题：一是平台对 AI 操控的识别能力有限，依赖人工审查成本高，效率低；二是面对海量信息流，部分平台采取“宽容算法”以保证用户活跃度，反而为操控行为提供生长土壤；三是平台普遍缺乏可溯源的透明机制，使得操控责任难以界定。

平台治理策略往往因商业利益驱动而有所保留，特别是在涉及敏感议题或重要舆情事件时，平台常陷于“引导—压制—放任”之间的策略摇摆，加剧公众的不信任感与监管困境 [2]。

4.3 跨国治理与法律滞后

舆论操控行为具有明显的跨平台、跨语言、跨国界特征，给传统的单一国家监管体系带来挑战。目前多数国家尚未建立针对此类行为的专门法规或国际合作机制。在 AI 技术更新迅猛的背景下，法律更新周期明显滞后，形成“技术先行—治理滞后”的制度落差。

例如，在海外社交平台上传播的假新闻或 AI 生成图像难以受中国现行网络安全法规约束；反之，国外平台对国内公众产生影响时，往往以“言论自由”规避内容审查责任。这种制度鸿沟限制了全球治理协同与执法效率 [3]。

4.4 公众识别能力与媒介素养不足

舆论操控行为之所以能迅速引爆和发酵，公众的信息判断能力是关键变量。在 AI 内容呈现越来越“真实”的背景下，部分用户缺乏对信息真实性的基本判断标准，易被情绪化话语带节奏、误导，甚至参与信息扩散。

尤其是年轻群体与中老年网民，分别面临“信息过载”和“辨别能力弱”的双重困境。缺乏基础媒介素养教育，也限制了社会整体对 AI 操控风险的识别与抵御能力。这种“技术外部性”反映出公众教育体系在新媒体时代的短板。

五、多元协同的治理对策探索

5.1 技术应对路径：溯源、识别与标记机制建设

针对 AI 生成内容在舆论操控中的隐蔽性与可扩展性，必须强化技术层面的主动识别与源头控制能力。一方面，推动开发更精准的 AI 内容识别算法，包括基于语义风格、语用特征和生成轨迹的分析模型，对可疑内容进行风险评级；另一方面，建立内容溯源机制，对所有生成式内容添加“机器生成”标签或水印标识，提升信息透明度。

此外，建议设立“人工智能内容风控接口”标准，推动平台间互通“内容溯源协议”和风险识别数据模型，实现跨平台的协同感知与响应能力，减少舆论操控的扩散通道。

5.2 法律法规完善：行为界定与责任划分

法律是治理 AI 舆论操控的基础屏障。目前我国已在《数据安全法》《网络安全法》《人工智能标准化白皮书》等文件中提出基本原则，但仍缺乏对生成式内容、算法推荐风险、虚假信息制造的专门规制。

建议尽快出台针对 AI 内容操控行为的细则性立法，明确“生成内容不标识”“算法诱导偏向”“组织操控账号矩阵”等行为的法律边界。同时，应完善责任追溯机制，细化平台、技术提供方、实施主体之间的责任划分，为追责与执法提供依据。

5.3 平台自律与监管协同：提升治理透明度

平台作为舆论传播的关键枢纽，应强化“第一责任人”意识，构建事前防控、事中识别、事后溯源的闭环机制。一是加强对 AI 生成内容的标注制度，二是设置“异常行为识别模型”监控操控性账号集群，三是设立独立“内容审计团队”加强平台治理透明度。

政府监管机构则应建立与平台的数据共享机制，推动建立“平台治理绩效通报制度”与“违规惩戒黑名单”，通过技术+制度手段倒逼平台责任落实。同时，鼓励第三方机构参与平台风险评估，形成政府、平台、社会三位一体的共治格局 [4]。

5.4 公众媒介素养提升：增强舆情识别与应对能力

公众作为舆论操控的直接接受者，其辨别力与媒介素养是防控风险的最后一道屏障。应通过教育体系、媒体普及与平台互动等多元方式，系统提升公民的信息甄别能力和 AI 内容意识。

建议推动高校、中学开设“人工智能与信息判断”课程，提升青年群体对 AI 操控手段的识别水平；鼓励媒体设置“AI 舆论科普”专栏，传播案例型知识；平台也应开发“反操控工具箱”或“虚假内容检测插件”，赋予用户自查能力，形成全民参与的抗操控网络。

5.5 国际经验借鉴与跨国监管联动机制构建

AI 舆论操控具有明显的跨境传播特征，仅靠单一国家治理难以奏效。当前，欧盟已在《人工智能法案》中提出对“高风险 AI 系统”进行分类监管，美国、加拿大等国也正推动 AI 生成内容标识和平台透明度法案。

我国应积极参与国际 AI 治理规则的协商，推动构建“跨境信息识别共享平台”与“假新闻联合清理机制”，通过外交、平台合作与多边协定实现国际舆论操控的联合打击。同时，可设立“全球 AI 内容风险数据库”，提升对新型操控模式的预警能力。

六、结语：构建 AI 语境下的舆论治理体系

6.1 本文研究结论概述

本文围绕“人工智能操控舆论的风险分析与治理对策”主题，系统梳理了生成式 AI 在内容制造、情绪引导、群体操控中的主要手段与技术路径，指出其在舆论引导中所带来的信息扭曲、认知干扰及治理难题。通过案例分析，进一步揭示了 AI 操控行为在社交平台上的典型表现形态和传播机制，呈现出高隐蔽性、强组织性和传播裂变性等特点。

在此基础上，论文从技术、法律、平台、自律与国际协同五个维度提出了多元协同的治理策略，强调需构建以“风险识别—行为约束—社会动员”为核心的立体化舆论治理框架，以应对人工智能语境下日益复杂的传播生态。

6.2 对公共舆论安全的治理启示

本研究表明，人工智能已成为影响舆论格局的关键变量，其风险不仅在于虚假内容本身，更在于操控路径与信息生态的系统性重塑。因此，治理策略不应仅停留于技术补漏层面，而应转向系统性结构优化：包括算法透明、内容可溯源、平台问责机制与公众媒介素养提升等。尤其在热点事件、选举舆情、突发公共事件中，AI 操控易引发情绪极化、信息偏置甚至社会撕裂，治理工作应前置部署、快速响应，并引导公众理性参与公共讨论，维护信息空间的真实性与信任度。

6.3 研究的不足与后续探索方向

本研究仍存在若干限制。首先，案例研究以定性分析为主，未来可结合大数据挖掘、网络结构分析等手段，对 AI 舆论操控行为进行量化追踪与模式识别；其次，本文所涉及的平台与区域以中国与部分西方国家为主，尚未涵盖全球视野下更广泛的操控机制异同；此外，对于生成式 AI 的技术底层逻辑与算法黑箱问题，尚缺乏深入探讨。

未来研究可进一步聚焦多模态 AI 操控内容的检测机制、算法治理中的伦理责任划分、以及跨平台协同治理机制设计等方向，为全球数字舆论安全构建提供理论与实证支持。

参考文献：

[1] 张涛甫．人工智能推动舆论生态转型及其治理进路 [J]. 学术月刊，2024(2):34-42.

[2] 陶贤都，李肖楠．算法推荐影响社会舆论安全的风险及治理 [J]. 信息生态，2022.

[3] 王瑾．算法推荐下的网络舆情法治化建设 [J]. 法学，2023, 11(2):778-782.

[4] 胡晶晶，吴思佳．人工智能时代主流媒体网络舆论引导研究 [J].

[5] 生成式人工智能重塑舆论传播机制：特性、风险与规制 [J]. 信息技术与管理应用，2024,3(6):1-10.

法律·科技·社会

LAW
TECHNOLOGY
SOCIETY

3058-1742