情报理论与实践
Information Studies:Theory & Application
ISSN 1000-7490,CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：生成式人工智能作用下网络群体极化：形成机制、影响、治理
作者：王晰巍，邱程程，吴彦婷
网络首发日期：2025-08-07
引用格式：王晰巍，邱程程，吴彦婷．生成式人工智能作用下网络群体极化：形成机制、影响、治理[J/OL]．情报理论与实践.
https://link.cnki.net/urlid/11.1762.G3.20250807.1440.002

●王晰巍[1,2,3]，邱程程[4]，吴彦婷[4]
（1.吉林大学商学与管理学院，吉林　长春　130015；2.吉林大学大数据管理研究中心，吉林　长春　130015；3.吉林大学网络空间治理研究中心，吉林　长春　130015；4.吉林大学国家发展与安全研究院，吉林　长春　130015）

生成式人工智能作用下网络群体极化：形成机制、影响、治理*

摘　要：［目的/意义］生成式人工智能（GAI）作用下的社交媒体对当前社交网络信息生态产生了重要影响，GAI技术在社交网络中主体性的凸显，作为具有认知影响力的行动者直接参与群体意见的形成过程，并重构了网络舆论场信息生态格局。［方法/过程］基于大量文献，围绕GAI作用下的形成机制、社会影响和治理路径等三个科学问题展开系统性分析。［结果/结论］在理论层面，构建了“技术驱动—中观影响—宏观结构”的三元整合分析框架，解构了认知调节、社会情感、技术增强、网络重构四大形成机制。在机制深化层面，提出GAI技术异化的技术武装化、空间脱域化、系统失控化的三重嬗变机制。在治理创新层面，提出了生成式人工智能作用下网络群体极化的治理路径，创建了技术治理、制度创新、能力建设三维治理框架。研究有助于更好地理解GAI驱动下的网络群体极化的形成机制，为GAI驱动下的社交媒体网络群体极化风险的引导和治理提供理论和实践指导。

关键词：群体极化；生成式人工智能；形成机制；社会影响；治理路径

Network Group Polarization under the Action of Generative Artificial Intelligence: Formation Mechanism, Influence and Governance

Wang Xiwei[1,2,3], Qiu Chengcheng[4], Wu Yanting[4]
(1. School of Business and Management, Jilin University, Jilin Changchun 130015; 2. Research Centre for Big Data Management, Jilin University, Jilin Changchun 130015; 3. Research Centre for Cyberspace Governance, Jilin University, Jilin Changchun 130015; 4. Institute of National Development and Security Studies, Jilin University, Jilin Changchun 130015)

Abstract: [Purpose/significance] Social media influenced by GAI has exerted a profound impact on the current information ecology of social networks. By emphasizing subjectivity within social networks, GAI technology actively engages as a cognitively influential actor in shaping group opinions, thereby reshaping the information ecology of the online public opinion landscape. [Method/process] Drawing from a vast array of literature, a systematic analysis is conducted focusing on three scientific questions: the formation mechanism, social impact, and governance pathways under the influence of GAI. [Result/conclusion] At the theoretical level, it constructs a three-dimensional integrated analysis framework of technology-driven, meso-influence and macro-structure, and deconstructs four formation mechanisms: cognitive adjustment, social emotion, technology enhancement and network reconstruction. At the level of mechanism deepening, a triple evolution mechanism for GAI technology alienation is proposed, encompassing technological militarization, spatial disembedding, and systemic deregulation. At the level of governance innovation, a governance pathway for network group polarization influenced by generative artificial intelligence is introduced, alongside a three-dimensional governance framework of technological governance, institutional innovation, and capacity building. This study enhances our understanding of the formation mechanism of network group polarization driven by GAI, offering both theoretical and practical guidance for mitigating and governing the risks associated with social media network group polarization driven by GAI.

Keywords: group polarization; generative AI; formation mechanisms; social impacts; governance path

0 引言

社交媒体平台作为网络舆情的核心策源地，通过智能推荐算法加剧了回音室效应[1]与信息茧房的形成[2]，导致网络舆论场在典型话题影响时出现群体极化。这种技术与社会互动催生的回音室效应在过滤气泡持续作用下导致的网络群体极化，正持续引发学界和舆情监管部门的关注[3]。值得关注的是，生成式人工智能（Generative Artificial Intelligence, GAI）技术的突破性发展正在重构网络舆情这一演化进程。以生成对抗网络（GAN）和大语言模型

*本文为教育部人文社会科学研究一般项目“信息生态视角下人工智能生成内容人机交互风险及治理路径研究”（项目编号：24YJA870012），吉林省自然科学基金面上项目“重大突发事件下智能推荐算法对网络舆情演化影响及风险预警研究”（项目编号：20240101372JC）和2024年度吉林大学国家发展与安全研究院专项课题“人工智能内容生成的安全风险及治理路径”（GAY2024ZXW01）的成果。

（Large Language Models，LLMs，以下简称大模型）为代表的技术范式正在深刻影响网络舆情生态。GAI 技术在社交网络中主体性的凸显，使其不再仅是信息传播的中介工具，而是作为具有认知影响力的行动者直接参与群体意见的形成过程，重构了网络舆论场信息生态格局。

现有网络群体极化的研究成果呈现多维透视的研究格局。在理论建构方面，网络用户群体极化的研究集中在群体极化识别和群体极化形成机理，主要包括舆论[4]和虚假信息传播[5]等。随着计算社会科学范式的兴起，基于 Agent 建模的社会仿真[6]和基于深度学习[7]的观点动力学分析逐渐成为解析网络群体极化的新兴方法，但在跨学科理论融合方面仍存在解释框架碎片化的问题。在社会影响评估方面，社会学研究证实群体极化催化网络暴力[8]和网络欺凌[9]，而传播学研究则解释算法推荐系统通过过滤气泡加剧认知偏差的中间机制。在网络舆论场多学科交叉研究方面，每个学科领域都有自己独特的问题定义、术语、解释模型和方法论。这些虽然深化了传统社交媒体极化现象的理解，却普遍缺乏生成式人工智能作用下对网络群体极化的影响这一新兴技术的考量。从研究趋势和前沿进展看，呈现“三维跃迁”特征：一是方法论向计算范式跃迁，传统定性分析方法正被深度学习方法和社会仿真技术所迭代；二是影响评估向多级传导深化，形成“个体行为—群体互动—社会结构”的全链条评估；三是理论建构向危机—技术协同聚焦，尤其重视生成式人工智能 GAI 下颠覆性作用，可能重构极化形成的底层逻辑。

当网络群体极化的研究客体扩展到具有语义生成能力的 GAI 对网络极化的影响时，传统分析框架面临双重挑战：在技术层面，生成模型的内容涌现特性使得观点传播突破人工操纵的确定性边界。在认知层面，人类用户与 AI 的交互式对话正在重构社交网络信息传播中用户的认知机制。基于此，本文试图通过研究回答以下三个方面问题：一是 GAI 作用下网络群体极化形成机制是什么？二是GAI作用下网络群体极化会带来哪些社会影响？三是GAI作用下网络群体极化的治理路径是什么？本文的创新主要是解析 GAI 作用下网络群体极化的形成机制，揭示 GAI 技术介入下网络群体极化的社会影响及其呈现的新特征，提出 GAI 作用下网络群体极化的治理路径。在理论上，为相关学科更好地理解和分析 GAI 作用下群体极化形成机制、社会影响和治理路径，提供理论支撑。在实践上，为政府和社交媒体平台在参与网络舆论引导和政策干预时提供有效参考，提高政府解决 GAI 作用下的群体极化社会问题提供科学决策支撑。

1 生成式人工智能作用下网络群体极化的形成机制

GAI 作用下网络群体极化现象呈现出复杂的形成机制，可解构为 4 个相互作用的形成机制维度，构成“技术驱动—中观影响—宏观结构”的多层次形成机制框架，如图 1 所示。其中，技术驱动层，以 GAI 的核心能力构成网络群体极化现象的技术驱动源头，这些技术特性直接作用于个体认知与信息接触环境，是形成机制框架的微观基础；中观影响层，技术驱动直接导致中观层面的群体行为异化，表现为群体内部观点加速固化与外部群体间对立加深，是形成机制对群体行为产生的社会影响；宏观结构层，中观群体的行为模式汇聚形成宏观社会结构变化，表现为 GAI 作用下网络群体极化现象形成机制带来的社会结果。

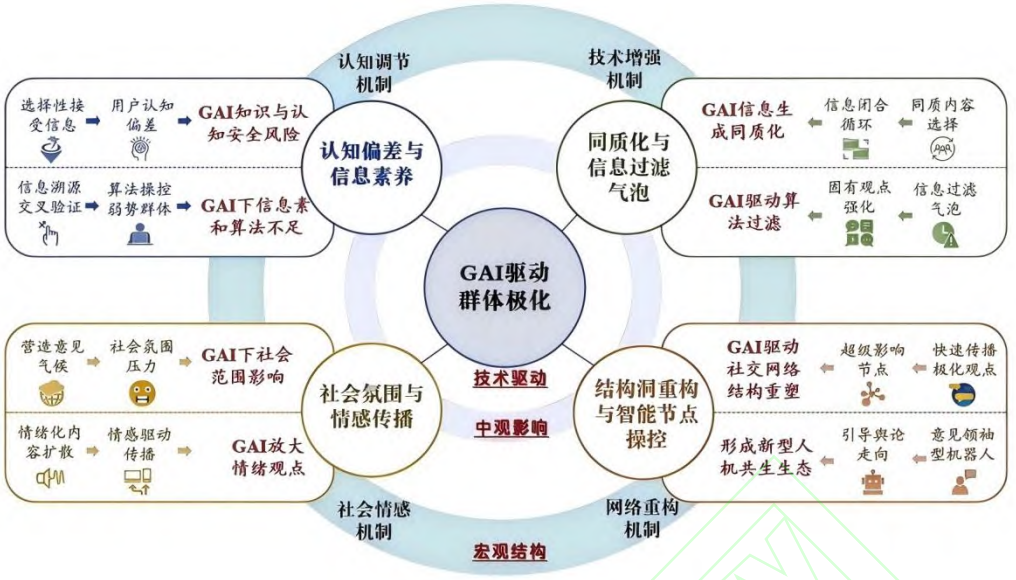

图 1 GAI 作用下的网络群体极化的形成机制框架

Fig.1 Formation mechanism of network group polarisation under the influence of GAI

1.1 认知调节机制：认知偏差与信息素养的双向调节

在社交媒体网络舆情演化过程中，认知偏差与信息素养的相互作用形成了群体极化的双向认知调节机制。研究表明，网民群体的认知偏差会对群体极化产生影响[10]。在社交互动层面，用户倾向于选择观点相近的互动对象形成社会选择偏差。在内容消费层面，算法推荐系统与用户主动搜索行为共同强化内容选择偏差，最终共同导致个人认知层面的偏差[11]，这种双重过滤机制导致个体认知系统形成信息茧房并导致认知安全风险，从而加剧了社交媒体的群体极化[12]。信息素养对网络群体极化也发挥着关键调节作用。教育鸿沟导致的社交媒体用户信息处理能力分化为两个维度：一方面，低教育群体普遍缺乏信息溯源与信息交叉验证能力，社会身份认同对认知判断的干扰效应在弱势群体中更为显著，其舆情的立场受社交网络中情感共鸣而非事实判断主导；另一方面，数字技术接受度的代际差异使得弱势群体面临更高的算法操控风险，这种结构性失衡导致网络信息传播演化呈现认知断层特征，即不同群体基于差异化信息素养形成对立观点，推动社交网络走向群体极化。

GAI 通过语义操控技术重构网络用户认知安全边界。大模型的对话式交互特征带来了潜在的知识安全和认知安全风险问题[13]。传统的信息素养教育关于媒介批判的能力，在 GAI 语境下升级为算法素养。在网络舆情事件的发展过程中，这种认知偏差使得人们更容易受到表面信息的影响，难以形成全面和客观的看法，可能在对社会现象进行评价时表现出过激的行为。随着人工智能生成内容（AI-Generated Content，AIGC）技术的应用和推广，全民数字素养教育进入智能时代[14]，需要具备信息素养的同时更要具有算法素养。社交媒体中的一些用户由于信息和算法素养的缺失，使得网络群体极化可能性大幅增加，同时其演化过程也变得越发难以掌控。

1.2 社会情感机制：社会氛围与情感传播协同放大

社会氛围与情感传播影响构建形成作用机制。前者促使个体通过观点自我披露获得社会认同。人们倾向于维持自己的认知、信念和态度之间的一致性。当接触到与现有观点不一致的信息时，可能会产生认知失调，个体会试图通过改变观点、忽略信息或合理化来减少这种不一致[15]。当人们参与社交网络时，对于网络舆情事件会产生自己的观点，当其发现社交网络中存在与自己观点相似的群体，就会学习他人相似观点中的信息，从而加深对自己所持有观点的认同，在社会氛围影响下加剧群体极化。情感传播加剧了社交媒体群体极化进程。社交媒体上的信息传播往往伴随着强烈的情感色彩，情绪在用户之间传播容易形成情感共鸣。社交媒体的情感传播存在明显的负面偏好，当内容标题唤起更强烈的愤怒情感时，用户分享意愿提升超过两倍[16]。研究表明推文中每增加一个情绪词，其被转发率将平均提升 20%[17]。

GAI 驱动的社交机器人能够通过“多层协同传播网络”，以社群与技术驱动形式营造意

见氛围，从而加深个体对自身观点的认同[18]。研究表明，仅需要占人类用户总数约 2%～4%的社交机器人就可以大概率（80%）操控舆论氛围，即使在低网络密度或低重连概率的情况下，也仅需要约 5%到 9%的社交机器人[19]。GAI 作用下社交媒体算法常常对内容的吸引力进行优化，使得极端情绪的内容更易获得关注和传播。同时，GAI 生成内容在负面舆情的萌芽、爆发、复燃与衰退各阶段均显著助推了谣言扩散和情绪激化，尤其在放大公众的不信任情感和加剧情感波动方面具有影响显著[20]。

1.3 技术增强机制：同质化与过滤气泡技术迭代

网络群体极化形成中，同质化选择与算法过滤协同作用构成了关键性催化系统。基于社会网络理论，同质化选择包含三个维度，即属性同质化（年龄、教育等人口特征趋同）、价值同质化（意识形态与价值取向匹配）及行为同质化（消费模式相似）。这种同质化选择倾向导致相似观点用户聚集，多维趋同促使社交网络形成结构性缺失的封闭社群，并导致信息的局限化，加深了群体内信息闭合循环引发的群体极化[21]。过滤气泡作为技术增强型同质化载体，通过数据分析技术形成用户动态画像实现精准认知操控。在信息过滤气泡中，用户只接触到与自己观点一致或相似的信息，相似观点得到不断强化，气泡内的用户群体开始盲目认同这些相似观点，并抵制其他观点的进入。久而久之，信息过滤气泡内的信息质量和可信度就会显著下降，在引发群体极化的同时也造成信息失真和情感极化。

GAI 作用下算法在社交平台的作用机制会逐渐深入，极大程度地改变了社交媒体中信息传播的生态环境[22]。用户看到的内容是通过人工智能驱动算法过滤的，这种算法强化了用户现有的信仰和偏好，可能会排除相反或不同的观点[23]。在长期接收同质化信息的情况下，网民群体容易陷入信息茧房并逐渐适应相对封闭的信息环境，相似的观点被不断重复和加强，使得群体成员的观点变得更加极端化，并产生回音室效应。

1.4 网络重构机制：结构洞重构与智能节点操控

复杂网络理论揭示社交媒体具有双重结构特征。小世界网络特性（平均路径长度端、聚类系数高）确保信息传播速度比随机网络快，而无标度无网络的幂律分布特性导致信息传播呈现显著异质性。这种结构使得信息可以迅速在网络中传播，但也可能加剧群体极化。许多社交网络还呈现出无标度网络特征，即网络中少数节点拥有大量的连接成为枢纽节点，而大多数节点的连接数较少，这种不均匀的连接分布可能导致信息传播的不均衡，使得某些观点或信息更容易被放大和传播。基于大模型的智能体通过“人机渗透—社群重构—网络重塑”的三阶段机制，推动社交网络从无标度解构向“超级节点主导型”结构演化，使社交平台逐渐形成人机混合交互态势，改变了现有的社交网络传播结构和信息传播模式[24]。GAI 及大模型的应用提供了与大量社交媒体用户建立“一对一”交互关系的机会，推动了社交网络结构呈现“再中心化”，出现新的社交网络超级节点[25]。

GAI 作用下智能节点通过三重机制强化极化效应。智能节点中社交机器人成为社交网络中新兴议程设置者[26]，意见领袖型社交机器人能够依托其广泛的连接和高频率的互动，快速将信息传播到社交网络各个节点，这种扩散效应可以在短时间内形成强大的舆论压力。在社交媒体传播信息过程中，GAI 作用下的社交机器人可能会被用于根据特定偏好和利益实现选择性传播，算法能够将边缘或激进观点视为主流观点，并将人们的注意力转移到激进内容上，从而助长群体极化现象[27]。当前，人机混合网络的动态博弈已形成人机共生的新型网络生态，这种社交网络的结构重构与信息生态共同演化，正在重塑数字社会的认知使群体极化加剧。

2 生成式人工智能作用下网络群体极化的社会影响

在对 GAI 作用下网络群体极化形成机制分析基础上，提出 GAI 技术异化下带来的“技术武装化—空间脱域化—系统失控化”三重嬗变社会影响。

2.1 技术赋能：虚假信息与谣言传播嬗变

社交媒体场域下网络群体极化现象通过算法增强的信息茧房效应，形成了极化信息的加速传播机制。基于深度学习的算法机制放大了传统社交媒体中的确认偏差效应，用户倾向于寻找、解释和记忆信息以证实自己的先入之见，这种确认偏误使得与个人信念相符的信息更容易被接受和传播形成认知闭环，使得与极化观点相契合的信息获得优先传播权。在此过程中，虚假信息与谣言[28]因契合群体价值取向而获得传播豁免，进而在极化群体内部加速传播。

当前 GAI 通过深度伪造内容生成、智能社交机器人集群、大型语言模型开发与部署、人工智能增强的社交机器人和微定向等方式大量生产、重构和传播虚假信息[29]。这种技术异化导致社交媒体舆论生态出现三重嬗变，信息载体因智能工具的介入操控人类认知呈现技术武

装化升级，传播主体因虚拟身份的泛滥形成空间脱域化特征，舆论客体因算法操控超越人类监管陷入系统失控化旋涡[30]。基于此，虚假信息和网络谣言传播带来了多重风险。微观层面，群体认知的持续极化导致社交平台上的理性对话空间被压缩，形成社会信任危机；中观层面，不同群体间的认知鸿沟可能演变为网络空间的价值对立，并通过线上线下互动向现实社会渗透；宏观层面，网络空间的意识形态安全面临 GAI 加持的新型挑战，并成为影响社会稳定的新型数字治理难题。因此，政府在 GAI 作用下的社交媒体舆情监管上面临双重困境。一方面，舆情监管者和用户对 AI 技术的迭代规律与传播机制存在理解滞后；另一方面，传统舆情监管手段难以应对 GAI 作用下社交机器人的动态规避策略。

2.2 认知重构：算法茧房与认知生态异化

在 GAI 的深度介入下，网络群体极化现象正催生新的社会传播危机。智能算法基于用户的历史行为轨迹、情感倾向、交互模式、兴趣和偏好，进而实现精准的个性化内容推送，这种机制虽然提升了用户体验，但也可能加剧信息茧房效应。基于深度学习的算法机制放大了传统社交媒体中的认知偏差，用户通过个性化推荐系统不断强化既有认知框架，形成算法茧房效应下的认知闭环。久而久之，多数个体停留在表层认知，不再主动辨别事件真伪，导致信息视野日益窄化[31]。这使网民从被动接收信息转变为主动吸纳极化观点，甚至丧失独立思考能力。

社交媒体用户长期处于信息环境的窄化状态，从被动接收信息逐步转化为主动接收极化观点。这种公众信息隔离会对社交网络发展产生负面影响，从而影响网络信息生态。有证据表明，大多数公众对 GAI 作用下的智能推荐算法认识不足，这使他们特别容易受到单一认知信息的影响[32]。网民群体长期处于信息隔离状态，不仅影响其信息获取的广度，还可能逐渐影响其自身的信息检索和表达能力，使得个体的批判性思维和辨别真伪能力逐渐衰退，出现认知封闭现象[33]。长期算法驯化会使社交媒体中用户认知发生变化，这种社交平台中用户认知重构将会引发公共话语解构、社会信任衰减和文化多样性危机。这种用户认知生态的变化已超越单纯技术问题，可能演变为数字文明时代面临的新挑战。

2.3 社会裂变：舆论极端化与复合性风险

GAI 作用下社交媒体正重塑社会共识的形成机制。网民群体高度依赖社交媒体获取时事热点、交流观点并参与社会决策。研究表明，持相同立场的用户在社交平台上互动频繁，在 GAI 推动下网络社群呈现认知集群化特征。如关于政治和环境等问题的讨论，支持者和反对者往往形成各自社交圈，并相互强化观点形成对立社群[34]。用户通过表情包、话题标签占领等新型对抗手段，在虚拟空间构建意识形态堡垒，导致舆论场域呈现两极分化态势[35]。社交媒体网络中部分网民群体极端化观点会转变为极端行为，表现为网络暴力、人肉搜索和恶意举报等，严重威胁网络空间生态，并可能导致社会分裂、极端行为和政治极化[36]。例如，某些国家在政治选举中，GAI 被滥用于定向传播虚假信息或煽动性内容，试图利用极端情绪来影响选民立场，进而干扰政治进程，这种技术异化现象可能加剧了政党间意识形态的对立。

GAI 作用下社交媒体平台通过信息过滤、算法推荐和情感传播等机制促进了群体极化，并在此基础上导致舆论极端化。普通网民与执政阶层之间的认知鸿沟可能因群体极化而扩大。这种由技术驱动的舆论极端化现象，正在深刻影响着社会稳定的根基和民主治理效能，构成数字时代亟待解决的复合性风险。

2.4 治理困境：认同危机性与规制悖论化

GAI 深度介入社交媒体舆情传播使新型风险样态不断变化。循环式反转舆情呈现周期性爆发特征，弥散性复合舆情加速跨平台扩散，框架化议题舆情持续重构公众认知，间断爆发式舆情形成脉冲式冲击波。这些新型舆情样态不仅加剧了受众间的认知对立与价值冲突[37]，更通过 GAI 作用下虚假信息事件不断模糊真相与假象边界，使公众陷入“后真相”认知困境。普遍性信息怀疑主义催生社会焦虑，系统性信任危机持续蔓延。用户对所见所闻产生普遍怀疑进而引发焦虑和不确定感，这可能降低对社会系统的信任[38]，更有甚者对网络社会乃至现实社会厌恶，否定情绪不断堆积，逐渐以悲观心态认知社会全貌[39]。

GAI 作用下算法推荐系统的动态适配性，使得信息茧房呈现智能化升级和多模态传播态势，加剧认知框架碎片化。作为社交平台内容的消费者和生产者，人们比以往任何时候都更容易策划和共同塑造个人社交圈和信息生态系统，然而人们可能会错误地将他们在社交媒体中看到的内容视为现实世界代表，加剧社交平台群体极化[40]。社交平台中舆情事件引发的群体极化，会造成情感对立现象又延伸到现实社会，在政治认同、社会分层、文化价值等领域

产生深度裂痕。这种数字鸿沟的演变可能导致社交媒体中公共话语空间持续萎缩，政策制定者可能面临共识凝聚难度倍增和社会包容性治理效能递减的双重困境。

3 生成式人工智能作用下网络群体极化的治理路径

针对 GAI 作用下网络群体极化带来的社会影响，提出技术治理、制度创新、能力建设三维治理框架。技术层面，优化 GAI 算法与社交平台环境，构建 AI 赋能的动态监测与预警系统。制度层面，推进跨平台联动协同治理，强化法律规制与构建伦理框架。能力层面，培育媒介素养与引导公众行为。

3.1 优化 GAI 算法与社交平台环境

一是构建透明化算法治理机制。建立 GAI 算法白箱化解释体系，通过开发开源算法框架和可解释性工具，强制披露内容生成的关键参数，开发有利于用户及时识别算法模型潜在错误或偏差的检测工具包，实现输入—运算—输出全流程可追溯，实时检测算法输出的情绪极化指数和观点集中度，提升算法决策的公平性和可信度。二是完善算法开发的全周期治理模型。在算法模型开发阶段，通过对抗性训练引入跨文化语料库，强制算法接触对立观点样本和异质化信息源，防止出现算法的自反式信息茧房现象[41]。在算法模型部署阶段建立人机协同治理模式和动态反馈修正机制，构建用户反馈、专家评估、舆情监测的三维校正系统。当监测到舆情话题的极化指数超过阈值时，触发观点补强机制，并向用户推送经过事实核查的异质化内容。三是重构价值导向推荐机制。采用公共价值加权算法，在传统点击率指标外增设信息质量系数、事实核查等级、观点多元指数等评估指标，实行兼具多元性与公共性的信息推荐策略。对于涉及重大公共利益的舆情话题，实施优质内容优先推送机制，优先推荐具有专业机构备案、多方信源验证的内容。同时开发情绪监测系统，对煽动性和极端内容进行分级标记，通过延迟推送等方式降低其传播势能。在推荐系统中嵌入认知缓冲内容模块，当社交平台持续接触同类观点时平台系统插入中立性信息，形成认知调节的减速带。四是开发智能化工具破解信息茧房。为公众提供便捷的信息去极化工具，平台可引入随机动态推送机制，增加偶遇信息的供给，促使用户跳出信息茧房，降低信息窄化的风险积累。此外，平台还可通过匿名化用户行为数据，降低算法模型对用户既有兴趣标签的依赖，防止算法刻板描绘用户画像，进而实现内容推荐的多元化。

3.2 构建 GAI 动态监测与预警系统

一是构建多模态极化信息实时感知系统。利用自然语言处理（Natural Language Processing,NLP）、情感分析、多模态识别、跨模态语义关联技术，构建文字、图像与音视频协同检测框架，精准识别社交媒体中具有煽动性、虚构性及语义歧义的极化内容，及时捕捉社交媒体平台存在的极端议题、对立情绪和极化社群。特别是在突发公共事件中部署自动化监测和分析工具，对特定关键词、话题标签及特殊用户社群进行重点跟踪。二是构建多层级的极化信息预警系统。实现与群体极化风险的有效干预和处置。设定极化风险指标阈值，如社交极化水平、负面情感指数等，评估特定议题在社交媒体平台中的极化风险等级，实现极化风险的精准量化与识别。基于用户节点级别的极化监测，支持更精细的极化用户识别与早期预警机制建设[42]，以实现对极化信息及其潜在扩散趋势的精准识别与干预。三是构建极化预警分级响应策略。低风险阶段部署 GAI 生成理性对话框架，通过语义重构技术自动生成反极化叙事内容（如事实核查信息、多元视角报道），以对冲极端言论的传播势能；中高风险阶段则启动人机协同治理协议，实施精准删帖、账号降权、信息矫正的组合干预策略，快速切断极端情绪与极化言论的传播链条。四是构建情绪极化与反转机制。引入反转调控策略（Reversal Control Strategy，RCS）/极化调控策略（polarization Control Strategy，PCS）算法模型，实现调控策略的互动性，将情绪极化的调节强度控制在合理阈值之内，实现精准治理和成本优化间的平衡[43]。同时，定期对系统参数进行调整和优化，构建覆盖多平台、多模态、多语言的多元异构数据池，提升社交平台群体极化信息动态监测系统的情境适应性。

3.3 推进跨平台的联动与协同治理

一是形成跨平台极化风险联防联控机制。构建生成式跨链溯源系统，依托 GAI 构建跨平台极化风险联防联控体系，通过深度伪造检测算法解析文本、图像、视频的生成特征，通过信息溯源技术手段识别群体极化信息的源头及其传播路径，并共享极化信息传播链。一旦发现其具有跨平台传播趋势，各平台可采取联动下架和限流等措施，实现平台间的协同治理。二是构建跨平台联防联控责任体系。通过签署多边治理协议明确各平台义务，构建全平台联

动的风险识别标准，建立涵盖文本、图像和视频的多模态风险特征图谱，通过区块链存证技术实现风险样本的加密流转与行为轨迹的可信追溯。实施平台治理效能信用评价制度，将联防联控响应时效和处置精度等指标纳入平台年度合规审计范畴，形成技术联防、责任共担、成果共享的协同治理模式。三是发挥跨平台算法意见领袖协作治理。在当前的跨社交平台传播生态中，许多算法意见领袖活跃于多个平台，使得算法意见领域的影响力不再局限于单一信息场，具有明显的跨平台迁移特性。开发多智能体协作系统，模拟跨平台舆论场的复杂互动，训练算法意见领袖智能体在异构环境中形成治理共识算法。四是构建意见领袖治理社会责任体系。平台可建立网络文明合伙人制度，对优质创作者实施分层认证管理。在身份标识方面，对通过社会责任评估的 KOL 授予首席网络文明官等梯度化荣誉标识。在算法激励方面，对主动参与辟谣协作、正能量话题引领的创作者给予流量加权推荐。在能力约束方面，将参与网络文明倡议和优质内容共建等正向行为转化为可兑换的创作权益。形成荣誉驱动、算法引导、能力培育的协同机制，推动意见领袖从流量追逐者转型为网络空间治理的共建者。

3.4 强化法律规制与构建伦理框架

一是完善技术应用的法律责任框架。加快制定和完善与社交网络 GAI 应用相关的专门法律，确立模型训练、生成内容、传播扩散全链条责任框架。明确 GAI 技术开发者承担的模型伦理审查义务，社交平台运营者履行算法透明度披露责任，信息发布者落实生成内容真实性验证要求。同时，通过用户协议、社群规范等文件将营造良好网络信息生态作为社交网络信息内容传播的倡导性规范[44]，呼唤构建良好舆论生态。二是不断健全 GAI 规制维度。建立深度伪造内容强制标识制度，要求在社交网络中传播的信息标注“AI 生成”水印及可信度系数；实施高风险算法模型备案审查机制，对具有舆论动员功能的算法模型实施穿透式监管。三是推进对算法黑箱的专项治理及推进制衡性监管。建立算法透明度标准和审查监督机制，要求平台定期以适当方式披露算法运行机制，接受第三方机构和公众的监督。引入算法审计制度，定期抽检平台信息推荐系统的价值偏离度；算法伦理专家委员会定期评估平台算法在信息多样性、价值中立性与社会公正性等方面的表现并进行修正。四是构建人机互嵌的敏捷型伦理治理框架。确立以人为本的价值导向，将人类伦理道德转化为可信的伦理代码和道德算法置入 GAI 内嵌式规范性结构中[45]，以有效维护公共价值和保障公共利益。增强用户交互透明度，向用户提供有关 GAI 生成内容的信息来源、准确性和生成方式等解释性线索，帮助用户更好地理解信息内容及其数据来源[46]，推动网络群体极化治理从外部规制向共治共享转变。

3.5 培育媒介素养与引导公众行为

一是建构认知、知识、能力联动的媒介素养培育框架。在认知维度，培育用户主动信息消费观念，通过认知干预引导其自身突破信息茧房，建立跨群体、跨观点、跨领域的信息接收机制，培养自身多元价值的对话思维模式。在知识维度，系统解析用户对社交媒体信息传播逻辑、GAI 作用下平台算法机制、极化信息形成机理与扩散路径的认知教育。在能力维度，提升公众对网络信息的批判性思维与虚假信息辨别能力、基于算法推荐的平台运作解析能力、对信息源头偏向性以及自身情绪与认知偏误的反思能力，帮助用户避免产生情绪化表达和非理性判断，塑造开放理性的舆论主体。二是构建公共理性引导机制。在情感连接层面，通过情感化叙事和场景化传播等具身传播手段，增强官方信息的情绪共振力，建立“政府—用户”情感共同体来消解极端情绪。在制度对话层面，搭建多层级对话平台，通过常态化议题征集构建对话、反馈、调整的动态响应机制，通过机器学习算法实现舆论引导的精准投放。在协同治理层面，创设多元共治的监督体系，组建由传播学者、技术伦理专家、平台开发者和用户代表构成的协同治理委员会，增强用户与政府间的情感连接与信任程度，平复用户极端情绪，缓解舆论信息极化趋势[47]。三是搭建兼具包容性与互动性的对话平台。鼓励来自不同群体的用户开展理性、友好的交流，政府主体借助对话平台快速了解公众意见分歧和情绪分布，有针对性地回应和引导，强化公众对治理过程的认同感，推动社会价值共识凝聚；吸纳多元主体参与对话平台的规则制定与运行监督，如邀请专家学者、平台开发者和普通用户等组成平台治理委员会，对平台秩序进行维护。

4 研究结论

本文的理论贡献：一是理论建构层面，分析了 GAI 作用下网络群体极化的形成机制。构建了“技术驱动—中观影响—宏观结构”的三元整合分析框架。通过解构认知调节、社会情感、技术增强、网络重构四大核心形成机制，系统揭示 GAI 技术介入下群体极化的形成机理。

二是机制深化层面，揭示了 GAI 技术深度介入网络群体极化的社会影响。提出 GAI 技术异化的技术武装化、空间脱域化、系统失控化的三重嬗变机制，为研判智能化社会系统性风险提供了创新分析框架。三是分析 GAI 作用下网络群体极化呈现的新特征。认知偏差与信息素养相互作用形成群体极化的双向调节，并通过语义技术重构网络用户认知安全边界，同质化选择与算法过滤协同作用构成了关键性催化系统，智能节点通过三重机制强化极化效应。四是治理创新层面，提出了 GAI 作用下网络群体极化的治理路径，创建了技术治理、制度创新、能力建设三维治理框架，技术层面优化 GAI 算法与社交平台环境，构建 AI 赋能的动态监测与预警系统。制度层面推进跨平台联动协同治理，强化法律规制与构建伦理框架。能力层面培育媒介素养与引导公众行为。

本文的实践价值：一是治理决策维度，构建基于三元整合分析框架的 GAI 协同治理框架。突破了传统群体极化治理的碎片化困境，构建起 GAI 作用下群体极化的多维度治理框架，从而提高政府解决社交网络群体极化社会问题决策的科学性，显著提升社交网络舆情群体极化风险的预见性与精准性。二是风险防控维度，评估网络群体极化对社会的影响。推动相关监管机构制定分阶段、有重点的舆情干预策略，提升网络舆情治理的前瞻性与有效性，为构建监测、预警、干预三级响应机制提供理论指导，推动舆情治理从被动响应向主动防御转变。三是社会修复维度，创建技术治理、制度创新、能力建设治理框架。可更好地化解群体极化中的矛盾，构建社会关系修复的数字化和社会化协同治理路径，促进不同观点之间的对话与沟通，实现社会矛盾化解与共识凝聚的双重治理目标，推动社会和谐。

本研究仅采用文献分析方法对 GAI 作用下群体极化的形成机制、社会影响、治理路径进行分析，没有结合定量分析的方法展开研究。未来将结合 GAI 作用下典型网络群体极化的现象展开定量及实证分析。□

参考文献

[1]XU Bo XU Zhengchuan, LI Dahui. Internet aggression in online communities: a contemporary deterrence perspective[J]. Information Systems Journal, 2016, 26(6): 641-667.

[2]ARSHAD S, KHURRAM S. Can government's presence on social media stimulate citizens' online political participation? Investigating the influence of transparency, trust, and responsiveness[J]. Government Information Quarterly, 2020, 37(3): 101486.

[3]XING Yunfei, WANG Xiwei, QIU Chengcheng, et al. Research on opinion polarization by big data analytics capabilities in online social networks[J]. Technology in Society, 2022, 68: 101902.

[4]HAN Xuehua, WANG Juanle, ZHANG Min, et al. Using social media to mine and analyze public opinion related to COVID-19 in China[J]. International Journal of Environmental Research and Public Health, 2020, 17(8): 2788.

[5]MEEL P, VISHWAKARMA D K. Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities[J]. Expert Systems with Applications, 2020, 153: 112986.

[6]YE Yuanjian, ZHANG Renjie, ZHAO Yiqing, et al. A novel public opinion polarization model based on BA network[J]. Systems, 2022, 10(2): 46.

[7]AJALA I, FEROZE S, EL BARACHI M, et al. Combining artificial intelligence and expert content analysis to explore radical views on twitter: Case study on far-right discourse[J]. Journal of Cleaner Production, 2022, 362: 132263.

[8]BLANCO E, FERNÁNDEZ-TORRES M J, CANO-GALINDO J. Disinformation and hate speech toward female sports journalists[J]. Profesional de la Información, 2022, 31(6): e310613.

[9]MARCIANO L, SCHULZ P J, CAMERINI A L. Cyberbullying perpetration and victimization in youth: a meta-analysis of longitudinal studies[J]. Journal of Computer-Mediated Communication, 2020, 25(2): 163-181.

[10]LEE J, KIM Y, KELSEY J P. Beyond wishful thinking during the COVID-19 pandemic:

How hope reduces the effects of death arousal on hostility toward outgroups among conservative and liberal media users for COVID-19 information[J]. Health Communication, 2022, 37(14): 1832-1841.
[11]KIM A, DENNIS A R. Says who? The effects of presentation format and source rating on fake news in social media[J]. Mis Quarterly, 2019, 43(3): 1025-1040.
[12]MODGIL S, SINGH R K, GUPTA S, et al. A confirmation bias view on social media induced polarisation during Covid-19[J]. Information Systems Frontiers, 2024, 26(2): 417-441.
[13]白云,李白杨,毛进,等.从知识困境到认知陷阱:生成式技术驱动型信息生态系统安全问题研究[J].信息资源管理学报,2024,14(1):13-21.（BAI Yun, LI Baiyang, MAO Jin, et al. From epistemological paradox to cognitive trap: research on security issues of generative technology-driven information ecosystem[J]. Journal of Information Resources Management,2024,14(1):13-21.）
[14]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. （ LI Baiyang, TANG Kun. Connotation change and countermeasures of digital literacy education for all in the context of AIGC[J]. Library & Information,2024,(3):32-39）
[15]MOQBEL M, KOCK N. Unveiling the dark side of social networking sites: personal and work-related consequences of social networking site addiction[J]. Information & Management, 2018, 55(1): 109-119.
[16]MCLOUGHLIN K L, BRADY W J. Human-algorithm interactions help explain the spread of misinformation[J]. Current Opinion in Psychology, 2024, 56: 101770.
[17]BRADY W J, WILLS J A, JOST J T, et al. Emotion shapes the diffusion of moralized content in social networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(28): 7313-7318.
[18]汤景泰,星辰.作为“武器”的谣言：基于计算宣传的认知操纵[J].新闻大学,2023,(8):16-30,116-117.（TANG Jingtai XING Chen. Disinformation as "Weapon": cognitive manipulation based on computational propaganda[J]. Journalism Research,2023,(8):16-30, 116-117.）
[19]张凌,刘琼,贺昌茂.基于沉默螺旋理论的社交机器人网络舆情干预研究[J/OL].图书情报知识 ,1-12[2025-04-12].http://kns.cnki.net/kcms/detail/42.1085.G2.20240927.1544.002.html.（ZHANG Ling, LIU Qiong, HE Changmao. The intervention of social bots on online public opinion based on the spiral of silence theory[J/OL]. Documentation, Information & Knowledge,1-12[2025-04-12].http://kns.cnki.net/kcms/detail/42.1085.G2.20240927.1544.002.html.）
[20]祁凯,周燕生.基于大语言模型生成内容的负面舆情态势恶化牵引作用研究[J/OL].情报杂志 ,1-10[2025-04-12].http://kns.cnki.net/kcms/detail/61.1167.G3.20250311.1110.002.html.（QI Kai, ZHOU Yansheng. Research on the Aggravating Effect of Negative Online Public Opinion Situation Based on Content Generated by the Large Language Model[J/OL]. Journal of Intelligence,1-10[2025-04-12].http://kns.cnki.net/kcms/detail/61.1167.G3.20250311.1110.002.html.）
[21]张玥,庄碧琛,李青宇,等.同质化困境:信息茧房概念解析与理论框架构建[J].中国图书馆学报,2023,49(3):107-122.（ZHANG Yue, ZHUANG Bichen, LI Qingyu, et al. Homogenization dilemma: concept analysis and theoretical framework construction of information cocoons[J]. Journal of Library Science in China,2023,49(3):107-122.）
[22]杨芳芳,宋雪雁,张伟民.国内信息茧房研究热点与演进趋势：兼论静态和动态双重视角

[J].情报科学,2024,42(5):169-176,185.（YANG Fangfang, SONG Xueyan, ZHANG Weimin. Research hotspots and evolution trends of domestic information cocoons: a dual perspective of static and dynamic perspectives[J]. Information Science,2024,42(5):169-176, 185.）
[23]SHIN D, JITKAJORNWANICH K. How algorithms promote self-radicalization: Audit of Tiktok's algorithm using a reverse engineering method[J]. Social Science Computer Review, 2024, 42(4): 1020-1040.
[24]张洪忠, 王競一. 社交机器人参与社交网络舆论建构的策略分析——基于机器行为学的研究视角[J]. 新闻与写作 , 2023(2): 35-42.（ZHANG Hongzhong, WANG Jingyi. Strategic analysis of social robots' participation in social network discourse construction: a study from the perspective of machine behavioural science[J]. News and Writing, 2023(2): 35-42.）
[25]张洪忠, 王彦博,任吴炯,等. 乌合之众的超级节点?AI大模型使用的人机网络结构分析[J]. 新闻界, 2023, (10): 12-19.（ZHANG Hongzhong, WANG Yanbo, REN Wujiong, et al. Supernodes of the crowd? human-machine network structure used by AI large models[J]. Journalism and Mass Communication, 2023, (10): 12-19.）
[26]ZHAO Bei, REN Wujiong, ZHU Yicheng, et al. Manufacturing conflict or advocating peace? A study of social bots agenda building in the Twitter discussion of the Russia-Ukraine war[J]. Journal of Information Technology & Politics, 2024, 21(2): 176-194.
[27]MCNEIL-WILLSON R, GERRAND V, SCRINZI F, et al. Polarisation, violent extremism and resilience in Europe today: an analytical framework[R]. BRaVE Project, 2019.
[28]ALKHODAIR S A, Ding S H H, Fung B C M, et al. Detecting breaking news rumors of emerging topics in social media[J]. Information Processing & Management, 2020, 57(2): 102018.
[29]胡泳. 人工智能驱动的虚假信息：现在与未来[J]. 南京社会科学, 2024, (1): 96-109.（HU Yong. AI-driven disinformation: present and future[J]. Nanjing Journal of Social Sciences, 2024, (1): 96-109.）
[30]张文祥.生成式人工智能虚假信息的舆论生态挑战与治理进路[J].山东大学学报(哲学社会科学版),2025,(1):155-164.（ZHANG Wenxiang. Governance approaches of public opinion: ecosystems and challenges from generative AI disinformation[J]. Journal of Shandong University(Philosophy and Social Sciences),2025,(1):155-164.）
[31]HE Yiqing, LIU Darong, GUO Ruitong, et al. Information cocoons on short video platforms and its influence on depression among the elderly: a moderated mediation model[J]. Psychology research and behavior management, 2023,(16): 2469-2480.
[32]CALICE M N, BAO Luye, FREILING I, et al. Polarized platforms? How partisanship shapes perceptions of "algorithmic news bias"[J]. New Media & Society, 2023, 25(11): 2833-2854.
[33]YUAN Xiaofang, WANG Chunyun. Research on the formation mechanism of information cocoon and individual differences among researchers based on information ecology theory[J]. Frontiers in Psychology, 2022, 13: 1055798.
[34]NEUBAUM G, CARGNINO M, MALESZKA J. How Facebook users experience political disagreements and make decisions about the political homogenization of their online network[J]. International Journal of Communication, 2021, 15: 20.
[35]WAKEFIELD R L, WAKEFIELD K. The antecedents and consequences of intergroup affective polarisation on social media[J]. Information Systems Journal, 2023, 33(3): 640-668.
[36]KENNEDY E H, MUZZERALL P. Morality, emotions, and the ideal environmentalist:

Toward a conceptual framework for understanding political polarization[J]. American Behavioral Scientist, 2022, 66(9): 1263-1285.
[37]张爱军,贾璐.类ChatGPT人工智能语境下网络舆情安全的风险样态及其规制[J].情报杂志,2023,42(12):180-187.（ZHANG Aijun, JIA Lu. Risk patterns and regulations of online public opinion security in the context of ChatGPT[J]. Journal of Intelligence,2023,42(12):180-187.）
[38]欧阳康,计效宇.人工智能背景下舆情复杂性及其智慧化应对[J].天津社会科学,2025,(1):44-50.（OUYANG Kang, JI Xiaoyu. Public opinion complexity and intelligent response in the context of artificial intelligence[J]. Tianjin Social Sciences,2025,(1):44-50.）
[39]GUPTA S, JAIN G, TIWARI A A. Polarised social media discourse during COVID-19 pandemic: evidence from YouTube[J]. Behaviour & Information Technology, 2023, 42(2): 227-248.
[40]WILSON A E, PARKER V A, FEINBERG M. Polarization in the contemporary political and media landscape[J]. Current Opinion in Behavioral Sciences, 2020, 34: 223-228.
[41]曹冬英.“自反式信息茧房”：生成式人工智能的构筑机理与超越[J].探索与争鸣,2025,(2):153-166,180-181.（CAO Dongying. “Reflexive Information Cocoon”: the construction mechanism and transcendence of generative artificial intelligence[J]. Exploration and Free Views,2025,(2):153-166, 180-181.）
[42]MADRAKI G, OTALA J, VAHIDPOUR B, et al. Polarized social media networks: a novel approach to quantify the polarization level of individual users[J]. Information, Communication & Society, 2025, 28(3): 361-395.
[43]TAO Yuqi, HU Bin, ZENG Zilin, et al. Detecting and regulating sentiment reversal and polarization in online communities[J]. Information Processing & Management, 2025, 62(1): 103965.
[44]侯东德,张丽萍.生成式人工智能背景下网络信息生态风险的法律规制[J].社会科学研究,2023,(6):93-104.（HOU Dongde, ZHANG Liping. Legal regulation of network information ecological risks under the background of generative artificial intelligence[J]. Social Science Research,2023,(6):93-104.）
[45]冉连,张薇.AIGC中的深度伪造信息:生成机理与治理策略——基于行动者网络理论的分析框架[J].信息资源管理学报,2025,15(2):137-150.（RAN Lian, ZHANG Wei. Deep fake information in AIGC: generation mechanisms and governance strategies: ananalytical framework based on actor-networktheory[J]. Journal of Information Resources Management,2025,15(2):137-150.）
[46]SHIN D, KOERBER A, LIM J S. Impact of misinformation from generative AI on user information processing: How people understand misinformation from generative AI[J]. New Media & Society, 2024: 14614448241234040.
[47]王雅梦,呼大永,祝宇琳,等.群体舆论和政府沟通对网络用户舆论极化的影响[J].管理科学,2024,37(2):54-68.（WANG Yameng, HU Dayong, ZHU Yulin, et al. The influence of public opinion and government communication on opinion and argument polarization of internet users[J]. Journal of Management Science,2024,37(2):54-68.）

作者简介：王晰巍，女，1975 年生，博士，教授，博士生导师。研究方向：网络舆情知识组织与用户信息行为。**邱程程**，女，1995 年生，博士。研究方向：群体极化。**吴彦婷**，女，2001 年生，硕士生。研究方向：信息迷雾。
作者贡献声明：王晰巍，提出研究命题、研究思路，修订论文。**邱程程**，论文撰写与修改。**吴彦婷**，协助资料收集与论文修改。
录用日期：2025-08-06