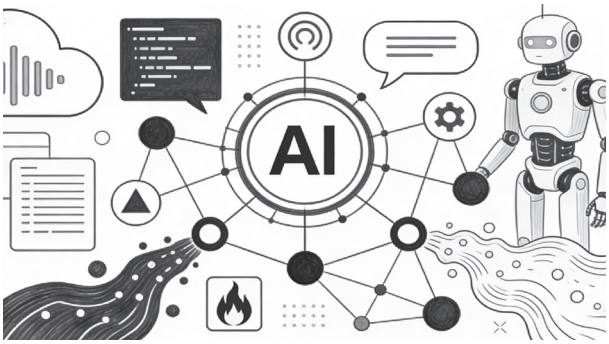
生成式人工智能可能引发的舆论风险及防范对策

文丨李思宇

人工智能技术正广泛应用于内容生成与传播、智能化分析决策、技术攻防等领域，在带来技术红利的同时，也通过数据投毒与污染、舆论操控和煽动、意识形态渗透等方式，对舆论安全造成冲击。因此，亟须探讨针对性的应对策略，以期为完善人工智能风险治理体系，推动人工智能安全、健康、有序发展提供理论参考。



（配图由 AI 生成）

截至 2025 年 4 月，我国人工智能专利申请量超 157 万件，占全球申请量超过 38%，居世界第一。人工智能技术正广泛应用于内容生成、信息传播、智能分析、辅助决策等领域，既赋能千行百业，也暴露出数据污染、算法偏见、舆情操纵等潜在风险，冲击舆论安全和网络生态防线。如何加强对人工智能技术跃迁可能引发舆论风险的防范和治理，推动算法向优、科技向善，成为优化人工智能风险治理体系的急切之问。

生成式人工智能可能引发的舆论风险

“数据投毒”和新型网络水军加剧谣言生成扩散，影响和干预舆论走向

一是“原生式谣言”难辨真伪，“数据投毒”误导公众认知。生成式人工智能通过对海量数据进行学习和分析，能够生成极具真实感、真伪难辨的内容。在图像生成领域，生成式人工智能模型能够模拟物理世界的各种元素，让缺乏专业辨别能力的普通用户信以为真。在语音生成领域，TTS 等技术能够精准捕捉人物的语调起伏、呼吸节奏和情绪波动，制作以假乱真的音频。视频理解与生成算法将图像与音频结合，能够完成唇型动态、微表情变化与语义内容的毫秒级匹配，生成高度逼真的视频片段。这种谣言突破了传统感知界限，对社会舆论的误导性极强。同时，虚假信息可能作为新的训练数据被再次“喂养”给学习模型，形成“垃圾进，垃圾出”的恶性循环。中国互联网联合辟谣平台对 2025 年网络谣言的分析显示，网络谣言主要集中在利用人工智能生成虚假信息、杜撰灾害事故及炒作民生热点等方面，产生混淆视听、误导认知等负面影响。

二是谣言生产批量生成，新型网络水军推动舆情发酵蔓延。人工智能降低了谣言生产的技术门槛和经济成本，能够批量快速生成结构完整、逻辑自洽的谣言。通过自动化脚本及多线程等技术的配合，新型网络水军已经从初始的机械性灌水回帖发展为高度类人化和仿真化的自主社交和虚拟发言，可以不间断地根据特定议题、基于固定立场在不同平台进行同质化的观点输出，以先占优势和规模效应操纵舆论声势。例如，2024 年江西南昌某 MCN 机构被公安机关查处。该机构构建的全自动谣言“生产线”，首先从国内某热点预测网站获取关键词，生成文章标题；再通过语音交互型 AI 工具抓取网络信息，生成虚假文本并匹配相关图片；最终借助自动化发布软件，将谣言批量上传至多平台账号对外发布。该机构最高峰时单日可生成 4000 ～ 7000 篇谣言，通过造谣引流、舆情敲诈，每日非法牟利超万元。另外，甘肃、浙江等地近年均发生利用人工智能编造传播虚假新闻的案件。此类犯罪行为成本低、危害性大，直接影响网络安全和行业发展。

算法偏见和信息茧房导致舆论短视，污染和冲击网络生态

一是算法偏见助推谣言弥散，网络秩序和社会信任遭到破坏。社交媒体平台和搜索引擎的算法推荐机制看似基于

用户浏览历史，精准推送可能感兴趣的内容，但事实上算法推荐正被经济利益、话题制造、运营策略等隐蔽操纵。算法推荐往往将注意力聚焦于冲突点和流量，将一些话题性强、能够吸引大量点击和互动的人工智能生成谣言视为优质内容推送给更多用户，成为谣言的放大器和发酵缸。根据中央网信办数据，网络谣言主要聚焦教育考试、交通出行、灾情事故等领域。不法分子借操纵算法推荐、编造虚假信息，达到宣传引流的目的，形成网络灰黑产业链。多个自媒体散布“2025 年高考答案流出”“考生高考作弊家长用钱摆平”等谣言，误导考生及家长。编造“贵州榕江洪灾造成 13 人死亡 2 人失联”“云南德宏遭遇严重洪灾”等虚假信息，炮制“河南驻马店物流港爆炸致 50 人死亡”“深圳地铁 11 号线爆炸”“四川九寨沟景区发生大巴坠崖事故”等谣言，引发社会性恐慌和公众误判。

二是信息茧房加重认知偏差和情绪极化，批判性思维和差异化选择受到限制。凭借“数字脸谱”迎合个体偏好的算法操控，因“越刷越精准”增强了信息茧房的回音壁效应。在高度同质化的信息链条中，长期与相似观点的人群进行互动，非常容易加深固有观念和刻板印象。部分资本掌握平台的内容生产、流量分发规则，并将其用于社会热点事件的议程设置、节点引爆、情绪操纵和节奏引导，导致不同群体之间的隔阂和冲突加深。例如，2025 年 5 月，“上海虹桥高铁站拦门事件”视频引发广泛关注。旅客被车门夹住的场景配以“车门夹人”“拦门闹剧”标签迅速传播。自媒体刻意强化“上海”“高铁冲突”等关键词，利用公众对不文明行为的痛感制造对立，瞬间点燃网民愤怒情绪。但根据铁路部门通报，该旅客系因突发身体不适匆忙下车，且经妥善处置并未影响列车正点发车。这场舆论风波反映了自媒体断章取义的流量狂欢、网民情绪化审判的惯性思维及平台算法对碎片化信息的推波助澜。信息茧房以标签化叙事替代理性追问、以情绪宣泄压倒事实核查、以虚假共识制造群体对立，导致“谣言跑在真相前面”。

技术霸权产生新型“数字殖民”，加剧意识形态渗透和政治认知操控

一是技术霸权成为认知颠覆的侵略工具，增加了意识形态安全治理难度。生成式人工智能正成为境外势力投毒虚假信息、窃取关键数据、挑拨意识形态、威胁他国安全的隐形武器。根据国家安全部发布的信息，某境外反华敌对势力通过深度伪造技术生成虚假视频，企图向境内传播误导性舆论、制造恐慌，对我国国家安全构成威胁；因开源库发生故障，某境外 AI 模型的部分用户可以看到其他活跃用户聊天记录中的名字、支付地址等隐私信息，如不注意防范，相关人员信息可能成为境外间谍情报机关实施拉拢策反、渗透破坏活动的指引，进而危害国家利益；个别西方国家和境外势力通过批量操纵账号，在社交媒体大肆发布特定观点或情绪化内容，妄图利用性别对立、劳资纠纷等议题进行渗透破坏、影响社会和谐稳定。

二是技术渗透操控国际舆论博弈，引发意识形态偏移风险。布鲁金斯学会评论文章《人工智能的政治：ChatGPT 和政治偏见》指出，ChatGPT 存在政治偏见。华盛顿大学、卡内基－梅隆大学和西安交通大学的研究发现，人工智能模型具有差异化的政治偏好。部分国家利用 AI 在敏感问题上输出具有特定指向的内容，为唱衰中国经济、制造认知分歧提供便利。部分 AI 呈现系统性偏向西方倾向，当使用英文就历史归属问题进行提问时，回答中存在明显的回避、淡化客观史实问题，甚至输出暗含错误信息或引发歧义的内容，意在通过大模型加剧历史分歧，解构国家安全根基。

防范生成式人工智能引发舆论风险的对策建议

清源流：建立信息溯源、全链路监测和联动处理机制，管控虚假信息生成源头

一是防范数据污染和 AI 幻觉，加固安全防护体系。一方面，确保训练数据客观准确、去污降噪，防止 AI 模型带毒输出。应以《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规为依据，完善数据分类分级管理制度，要求数据供给方、标注方、模型方落实数据清洗、敏感内容过滤义务，使用真实、准确、客观、多样且来源合法的训练数据，及时过滤失效、错误、偏见数据。研发机构应就训练数据进行共享和清理，构建数据白名单、算力集群和模型训练集群，以建设并行平台等方式加速数据净化，避免虚假、不良样本进入训练集。另一方面，加强生成合成内容标识治理，实现数据服务可审核、可监督、可追溯、可信赖。持续完善《人工智能生成合成内容标识办法》及《网络安全技术　人工智能生成合成内容标识方法》，构建面向人工智能全生命周期管理的技术安全治理体系。明确文本、图像、音频、视频等多模态的标识要求，细化显式、隐式标识的技术实现方式，全面覆盖制作源头、传播路径、分发渠道等关键环节，以技术痕迹可核查、内容来源可回溯，实现人工智能从生成到传播的全链条安全治理。

二是构建“识别－拦截－核查－辟谣”的闭环式防御体系，阻断网络谣言传播链路。一方面，压实平台主体责任，用技术穿透撕开拟真伪装。平台应升级信息审核体系，整合自然语言处理、图像水印分析、行为序列建模等技术，识别 AI 生成内容在句式一致性、图像光影异常点、文本低熵值等方面的隐性特征，通过账号登录 IP 集群、互动频率、语义相

似度等“行为指纹”定位 AI 水军账号池。建立“关键词 + 情感倾向 + 传播路径”多维监测模型，对短时间内爆发的相似内容、批量点赞评论等异常互动及时触发预警，标记高风险传播链。对涉政、民生等敏感领域的信息异常传播，启动“技术初筛 + 人工复核”双保险，对水军账号实施“发现即拦截 - 溯源即封禁 - 扩散即限流”管控。另一方面，提高“权威核查 + 官方通报”的舆论对冲力，形成“正声盖过杂音”的传播态势。构建集主流媒体、第三方核查机构、专业平台于一体的多层次、联动式辟谣网络，通过“权威鉴定 - 实时发布 - 多平台推送 - 全网覆盖”及时回应社会关切。辟谣时不仅针对具体事件，更要揭示人工智能进行舆论操纵的基本原理和技术手法，通过案例演示、互动问答等方式进行可视化呈现，帮助公众增强辨别能力。

阻狂澜：加强算法监管、破除信息茧房，优化网络舆论生态环境

一是构建“算法纠偏 - 责任锚定 - 监管约束”治理体系，破解算法黑箱。政府层面，应推动算法综合治理工作常态化和规范化，完善《互联网信息服务算法推荐管理规定》等法律法规，加强算法知情权、算法选择权、用户隐私权等权益保护。对社会关注的“大数据杀熟”“引诱沉迷”“引导舆论”“特殊群体保护”等问题进行专项整治，解决算法价值导向偏差、优质内容呈现不足、流量分配重指标轻质量等突出问题。落实算法备案机制，健全算法审查机制，对算法进行定期审查和评估。行业层面，企业应加强自律，提高算法透明度，主动进行自纠自查。重视用户意见反馈，及时调整优化算法模型，主动清理有害信息，处置违规账号，引入多样化推荐机制，以多元信息展示代替单一信源主导。建立内部审核机制，定期检查算法的公平性与规范性，推动算法向优、科技向善。

二是鼓励前置求证和辩证讨论，提升公众智能媒介素养和信息辨别能力。应将 AI 与新媒体素养教育纳入教育培训计划，重点提升公众对个人隐私的关注、对接收信息的识别与反思、对知识产权的重视、对网络安全法律法规的掌握等方面素养，从信息获取、源头甄别、批判思维、风险防范等维度加强宣传教育。通过专题讲座、模拟训练和社会案例分析，向公众普及查看内容标识、验证信息源等人工智能生成内容识别技巧，增强公众对谣言的识别和防御技能，引导公众从权威渠道获知真实信息，避免情绪型、跟风型的谣言扩散。拓宽人工智能安全投诉举报、核实验证、公开曝光渠道，营造识谣、辨谣、拒谣的全民认知生态。

固堤坝：构建技术筑防、舆论反攻的立体化防御体系，提高意识形态免疫能力

一是以技术升级制衡技术霸权，实现“源头发送端—境内传播链”双重阻断。应坚持正确的舆论导向，以数据人民性、对话针对性、内容思想性为指向，充实大模型训练优质语料内容。针对涉意识形态的不同场景进行差异化、分众化、层级化的分析研判，重点识别伪造经济增速、篡改失业率等经济唱衰论的虚假数据特征，以及极端化群体标签等政治分歧话术的情感操纵模式，以数据的敏感程度、规模体量、安全级别为依据，进行合法化、合规化、合理化筛选，避免数据违规利用、过度收集和不当泄露。同时，应加大多模态大模型等领域的研发投入，突破跨语言语义理解、深度伪造检测等核心技术，升级技术防御体系，筑牢意识形态风险防火墙。

二是丰富主流话语表达方式和叙事策略，拓展中国声音的传播空间和价值认同基础。应推动传统媒体、新媒体和社交媒体相互配合、优势互补，构建多媒体立体化传播矩阵，扩大正面叙事触达率。发挥主流媒体、旗舰媒体的带动效应，擦亮《央视快评》《国际锐评》《大湾区之声热评》等评论品牌，冲破外方针对中国的威胁性言论，解决海内外受众信息逆差问题；激发企业、社会组织和公众的参与活力，根据不同受众特点策划本土化传播、精准化传播、全球化传播路径。针对外部势力借 AI 歪曲事实、煽动群体对立、制造虚假共识等意识形态操控和颠覆主流价值观行为，引导民间团体通过 Twitter、Facebook 等海外社交媒体发起话题挑战，形成“主流媒体引航 + 大众传播渲染”互补效应，提升中国话语说服力和国际舆论引导力。

结束语

人工智能技术发展引发的虚假信息扩散、舆论极化等风险，既考验着对技术创新的边界把控，也对舆论治理的系统性、前瞻性提出了更高要求。未来需进一步强化技术向善的价值导向，以管控虚假信息生成源头、优化网络舆论生态环境、提高意识形态免疫能力为重点，将风险防范嵌入技术研发、应用和推广全流程，构建数字时代的舆论生态新格局。

作者简介： 李思宇　中共辽宁省委党校

责任编辑：金烨　投稿邮箱：zhouhl@staff.ccidnet.com