面对AI谣言,治理难题何解?广州市政协常委黄丽丽倡议——

让每一次转发都成为守护真实的行动

当下,AI谣言依托生成式人工智能技术的便捷性不断蔓延,不仅干扰公共信息生态,更对社会信任体系和稳定局面造成威胁。面对AI谣言的泛滥,我们该如何精准施策、有效治理?新快报记者就此问题采访了多位密切关注人工智能领域的广州市政协委员,以期获取更多的见解和建议。

■采写:新快报记者 陈慕媛 林钢威 ■摄影:新快报记者 龚吉林

现象

利用AI技术制作虚假场景

清华大学新闻与传播学院新媒体研究中心于2024年发布的报告《揭秘AI谣言:传播路径与治理策略全解析》指出,人工智能工具的不当使用,使得AI谣言量高速增长、逐步成势,近半年信息量增长65%。

新快报记者在中国互联网联合辟谣平台搜索关键词"AI",发现自2025年以来,该平台已发布多则AI谣言相关辟谣。

2025年11月,网民曾某通过AI软件生成一段"大黑熊出现在漓江边"的虚假视频,并将该视频发布至某短视频平台。有关部门表示,经核查,视频内容纯属虚构,漓江沿岸区域近期并未出现过黑熊活动的相关记录。曾某称其发布视频并非刻意博取流量,只是一时兴起利用AI技术制作了虚假场景视频并发布。视频发布后迅速在本地社交圈传播,引发部分市民对周边环境安全的不必要恐慌。

这并非个案。近期,张某伙同李某、韩某在短视频直播间为吸粉引流、获取打赏、网上带货,多次虚假演绎境外人员在国内建设实验基地,残害婴儿、绑架主播等剧本,配AI图片视频吸引眼球,引发直播间众人围观,扰乱网络空间秩序。

特征

AI谣言具有情绪优先性

在知微数据等联合发布的《AI谣言的特征、危害与综合治理路径》报告指出,从技术特征上看,AI谣言具有超真实伪造、个性化投喂以及低成本量产三大特征;从内容特征上看,AI谣言具有热点寄生性、情绪优先性以及圈层化叙事三大特征。

新快报记者留意到中国互联网联合辟谣平台对2025年2月网络谣言的梳理分析。网上数据监测和网民举报显示,当月网络谣言主要集中在利用AI技术生成虚假信息、杜撰灾害事故信息以及炒作社会民生热点等方面,混淆视听、误导认知,造成不良影响。

中央网信办公布了2025年"清朗"系列专项行动整治重点,包括整治"自媒体"发布不实信息、整治AI技术滥用乱象等八大重点,传递出保持利剑高悬,严厉惩治各类造谣传谣的决心。

委员谈

●广州市政协常委、广州市新联会副会长黄丽丽

让每一次转发都成为守护真实的行动

国家网信办等四部门联合发布的《人工智能生成合成内容标识办法》于2025年9月正式施行,促进人工智能健康发展,规范人工智能生成合成内容标识。根据新规,内容平台需对AI生成的文本、图片、音视频等内容强制添加显式标识及隐式元数据标识,确保内容可追溯性。

普通网民可通过哪些特征快速辨识AI谣言?来自经济界的广州市政协常委、广州市新联会副会长黄丽丽说,AI谣言往往遵循一些固定的"套路模板",关注信源、细节与逻辑,能有效识别AI谣言或不实信息。

一看信源,典型AI谣言的信源表述,多是"知情人士透露""网传消息""内部消息显示"等模糊话术,无明确作者、无正规信源。二看细节,AI谣言为增强可信度,常捏造精确到具体数字,却无法提供数据出处,有的AI生成视频中还会出现手指数量异常、口型与声音不同步等细节漏洞。三看逻辑,AI生成谣言常出现时间矛盾、地名错误、或者几个不实信息源的循环引用,存在逻辑瑕疵。

黄丽丽表示,网民要主动思考并交叉验证信息真实性,还要时刻警惕情绪诱导,避免被煽动性内容裹挟。更重要的是,大家要负责任地分享与表达,不急于传播未经验证的内容,让每一次转发都成为守护真实的行动。

●广州市政协委员、工业和信息化部电子第五研究所总工程师万举勇

主流大模型缺乏内置"事实校验模块"

"AI生成谣言的核心是概率性文本生成+缺乏事实约束,本质是模型仅关注语言规律,而非事实正确性。"来自科学技术界的广州市政协委员、工业和信息化部电子第五研究所总工程师万举勇从技术层面介绍了AI生成谣言的机制。

他介绍道,训练数据中存在事实性错误的样本是其中一个重要原因。若训练数据中本身包含大量虚假、片面、误导性信息,比如网络谣言、未经证实的八卦、带偏见的内容,模型会学习到这些信息的"语言模式"和"表述逻辑",从而导致生成与事实相悖的信息。

他表示,当前主流大模型没有内置"事实校验模块",仅关注文本的语法通顺性、语义连贯性,而非内容的事实正确性,模型无法区分"文本描述符合语言规律"和"文本描述符合客观事实"。因此对于训练数据中冲突的信息,就如同一事件的不同版本,模型会随机融合或选择"出现频率更高"的版本,而非"更接近事实"的版本。

●广州市政协委员、民革广州市委员会常委徐科飞

率先探索"真实性权重"的行业标准

广州市政协委员、民革广州市委员会常委徐科飞认为,AI开发者在训练模型时加入"真实性权重"非常有必要。简单来说,就是让AI在"讲故事"和"讲事实"之间分得更清楚,尤其是在涉及公共安全、医疗、财经等领域时,必须把"真实"放在第一位。

他说,做法可以很朴素。第一,数据要干净。训练AI时尽量多用权威、可信的资料,减少假消息混进去的机会。第二,模型要自带"查一查"。让AI在回答事实性问题时自动对照知识库或权威信息,避免它"想当然"。第三,风险要分级。写小说、写诗歌的AI可以更自由;但做新闻摘要、科普问答的AI,就必须严格遵守真实性要求,必要时加上提示或水印,让用户知道内容来源。

"广州是全国数字经济的重要城市,本地企业在AI领域实力强。"徐科飞建议广州可以率先探索"真实性权重"的行业标准,让技术创新和社会责任同步推进。既保护创作空间,也守住事实底线。

专家说

●广州市律师协会副会长,民革广州市委员会委员、社法委主任唐以明

诱导生成违法内容应承担主要侵权责任

"现行法律、行政法规并未对生成式人工智能服务有明确的追责标准及监管规定。"广州市律师协会副会长,民革广州市委员会委员、社法委主任唐以明也是一名广州市政协委员。他提到,当前有效的规范文件是在2023年7月10日,由国务院互联网信息办公室牵头,经国家发改委、教育部、科技部、工信部、公安部、国家广电总局共同同意发布的部门规章《生成式人工智能服务管理暂行办法》(简称"《暂行办法》")。

针对生成式人工智能服务提供者,《暂行办法》在第7-8条中对其在数据训练、数据标注中应履行的义务提出了要求,在第三章(第9-15条)中提出了服务的规范,但《暂行办法》并未明确相关行为的追责标准。

唐以明指出,使用者通过提示词诱导生成违法内容或主动去除标识的,应依法承担主要侵权责任;仅辅助创作且未去标识的,负有"合理注意义务",比如,要对AI生成的事实性内容进行核对。

他提到,就通过提示词诱导生成违法内容或主动去除标识之行为是否构成"直接侵权",网络安全法并未作出直接规定。在现行法律规范体系下,对此类行为是否构成直接侵权的认定,应严格依据民法典侵权责任编所确立的四要件归责标准进行审查判断。

若传播平台未及时处理已知谣言的,按民法典第1195-1197条承担连带责任;对明显未标识的AI内容未尽抽检义务,特别是对显而易见的未标识违法

AI内容的未采取必要措施的,(比如,短视频平台对高流量AI视频的审核),网信部门可依据《网络信息内容生态治理规定》第24条处以罚款。