AI幻觉挑战下新闻生产的安全性路径

摘要：随着生成式AI发展迅猛，“效率化”应用AI的逻辑与新闻行业对真实性、公共性的追求之间存在方向性差异，而技术变革所带来的行业焦虑则放大了AI幻觉的风险。本文围绕新闻生产流程，提出强调“人本位”的价值认同，细化“人机协同”使用逻辑，强化技术反制和法律法规的治理体系，以确保新闻媒体在AI时代的信息安全与可靠性。

关键词：AI幻觉　新闻生产　新闻伦理　舆论安全　信息生态

◎ 汪茂盛　张低琦

当前，生成式 AI 已被受众普遍接受，截至 2025 年 8 月，国内 AI 大模型应用的个人用户注册量已超 31 亿[①]。这种高效集中的智能化大模型正迅速重构舆论生态，各大媒体纷纷“卷”入这场“技术媒介”革命。早在 10 年前，大数据算法就以标签化推送机制开始渗透信息传播场域，生成式 AI 持续加速信息生态环境的复杂性。由于其对信息的“饥渴”性纳入生成过程“黑箱”，导致许多独立语境下的问答呈现错误信息或非匹配信息回应，让人看来 AI 在“一本正经地胡说八道”。而这类“胡说八道”正通过每天海量信息学习和大模型升级不断变得“真假难辨”，出现 AI 幻觉。其层出不穷无疑给新闻生产安全带来严峻挑战，引发新闻内容真实性、传播逆差等问题。

AI幻觉带来挑战

技术变革焦虑与新闻失实忧虑的双重压迫。新闻真实性是新闻的生命线。AI 幻觉动摇的是媒体行业的根基。反观现实：媒体应用 AI 技术已是大势所趋，那为什么存在认知与实践差距呢？

首先，生成式AI制作财经报表、体育消息等高度模式化的新闻成本较低。以新华社打造的自动撰稿机器人“快笔小新”为例，从 2016 年报道里约奥运会的 500 余篇消息简讯与战报零差错到 2024 年发展成多模块的写作机器人，2—3 秒即可成稿，较大地提升了新闻生产效率。

其次，媒体在数字时代的系统性变革存在迫切性选择。低成本高效率的生成式 AI 指向短期收益最大化，但也带来“同速革新”后的价值降低，表现为作品形式机械和模式同质化严重。比如，中央广播电视总台推出虚拟数字主播后，各大省市级媒体纷纷跟进，绝大部分媒体由于经费、资源限制，以依托第三方平台免费试用虚拟数字主播为主，扎堆推出视频节目，试用期结束便停止使用。这是潜意识对变革“任务性”完成的惯性所致，剩下部分付费媒体也因受众对虚拟数字主播失去兴趣、作品表达生硬等问题接连放弃。这类来自社会环境的变革压力并不会随着单项

技术的应用而消逝，反而会愈演愈烈，与此共生的是大模型学习速度提升后 AI 幻觉频发对新闻真实性的冲击，让媒体难以平衡技术变革焦虑与承担新闻失实责任的忧虑。

虚假信息与真实新闻传播效能逆差被放大。网络谣言等虚假信息的传播因内容本身具有高争议性、高关注度而能快速抢占舆论场。这种以谣言为中心的漩涡传播模式，扭曲和挤压了真实新闻的传播场域，甚至造成“逆火效应”[②]，反而加强受众对谣言的印象。

首先，制造虚假信息的方式变得更隐蔽并且影响难以消除。2024 年 4 月，上海警方发现一则谣言在网络大肆传播，顺藤摸瓜锁定了一个用 AI 洗稿新闻换取流量变现的团伙。他们造假方式简单，在购买的 AI 软件中输入“标题劲爆”“结尾争议”“情节曲折”等指令，即可全自动洗稿。如此生产的“博眼球”假消息一天可发布上千篇，而警方和法院则耗费了 10 个月才将该团伙绳之以法。但其发布的虚假信息还能在各大社交平台评论区被提及，证明即便从源头上删除虚假新闻，但其影响已形成，消除的时间成本极高。

其次，AI 介入后引发媒体信任危机。以传统新闻内容生产来看，采写编校核发，每个环节都是基于客观世界的信息采集和主观世界的信息验证的主客一体过程。但所有环节都需要耗费时间成本，生成式 AI 则反之，直接根据现有语料生成新闻。随着 AI 幻觉逐渐增多，受众也对媒体公信力产生怀疑，即便主流媒体都要求 AI 作品打上“显性”标识，但该规定仅能对艺术类配图、艺术类视频起到一定约束，对新闻生产过程中的文字撰稿等还无法有效约束。

此外，虚假新闻污染后的信息生态环境造成关注度掠夺惯性。AI 生成的虚假信息刷新受众接受度阈值下限。曾经人们对“三星堆挖出航母”这类一眼假的新闻嗤之以鼻，如今当 AI 低成本生成的海量虚假信息渗入传播场，一个虚假信息被反复推送给同一受众后，便有了“三人成虎”的心理预期，受众在这样的“情绪新闻”反复刺激下被驯化，不再满足于真实却不亮眼的信息，反而追求争议性新闻，进一步为 AI 幻觉提供了受众市场。

新闻生产探索安全性路径

坚守人本位的认知前提。AI 技术的不断迭代，导致一部分人认为新闻记者这一职业将被 AI 取代。但从当下 AI 生成新闻中可以窥见，“AI 味”一时还难以消除，尤其是在传统媒体更为擅长的深度报道中，情感共鸣、批判思维、逻辑框架的构成与解构方面 AI 甚至起到反作用。媒体要从以下几个方面认识到人本位的重要性，理性地看待人机协同、人机共生关系，消除“技术在前认知在后”的本领恐慌。

首先，人的自然属性的“可达性”是 AI 不可替代的。人以客观物理形式存在于社会，可以面对面、点对点地去到新闻现场，与新闻人物进行接触，而这是 AI 从存在形态上无法办到的事情。这也从侧面反映了记者践行“四力”是以脚力为基础的观点。眼力、脑力、笔力归根结底都来自脚力，现场带给记者的是生理与心理的双重感受，是一个由新闻人物、新闻事件与当时所处环境构成的复杂系统，所引起的记者的思考与判断是 AI 无法模拟的。

其次，人的情感属性的“共鸣性”是 AI 不可替代的。受众为什么能识别出“AI 味”，因为 AI 表达是基于用户指令下的绝对理性。记者在生产稿件过程中，带有批判色彩的隐喻表达，逻辑的环环相扣和情感叙事的相互穿插，从本意到象征的重塑与延伸，这些都是记者作为人所具有的独特思维能力，也是受众作为人所共有的特征。比如，哔哩哔哩“食贫道”栏目推出的纪录片《卧底 30 天，我们亲身记录了韩国邪教里的层层套路》《勇闯南美毒窝！我们拍到了可卡因制作的全流程》，屡次登上微博热搜榜，足以证明这类依靠“蹲”出来的“深度思考”作品并不缺乏受众。

人机协同的坚持使用逻辑。降低深度报道中 AI 参与度。从机制来看，AI 创作是基于已有语料的二次加工，而深度报道则是一手素材的高创造性工作。深度报道中采访对象或者新闻事件有可能是首次面向公众公开，也就是说互联网环境缺乏采访对象或新闻事件的事实细节，比如采访对象身份和事件经过等。如果在此类报道生产中采取 AI 创作，AI 将对比模拟其他同平台、

同方向的深度报道，进行“算法”补全生成，无法将深度报道中记者在特定场景下感知到的细微瞬间进行数字标识或量化。在AI“黑箱”生成的过程中所产生的细节，很可能是语料的杂糅、拼接，造成AI幻觉。

从流程上看，深度报道运用AI的效率较低。撰稿过程中，因AI缺乏对具体新闻人物、事件的语料学习，仍然需要记者手动输入采访获取的一手资料先“喂”给AI，而其得出的结果也可能掺杂着高度隐蔽的失实细节，还需记者耗费时间核实AI结果的真实性，反而降低效率。

坚持内容把关“先机后人”原则。AI用于基础校对已非新鲜事，但AI对作品内容的把关必须前移。以字节跳动为例，作为国内最早将大数据算法推荐模式用于媒体领域的公司，一直走在全国媒体数字化技术应用的前列，其在内容把关上依旧保持“先机后人”审核流程。AI机器检测用于内容审核的第一步，形成了从数字审校到人工基础审校和人工疑难问题超高审校的内容把关团队。这种模式的形成并非偶然，是服务平台在保障安全性前提下经实践检验后采取的稳妥方案。AI不只对独立语境下的反讽、隐喻等文字含义无法识别，对于图片和视频的识别更为糟糕，至今没有一款AI软件能够精准识别出音视频作品的基础差错，更不用谈对影像化表达中的意向性解读。

AI监督手段呼唤数字化与行政化

技术反制有待推出。首先，需推出行业性的AI检测系统。现阶段，媒体生产对于AI的使用标识主要取决于记者的主动性与审核人员的判断力，图片视频相对而言更容易作出“是不是AI生成作品”的判断，但对于长篇文字，仅依靠人工很难判断是否存在AI介入痕迹，急需引入技术监督。因新闻机构的级别、资源各不相同，每个机构都推出自我研发的AI检测技术并不现实。即便每个媒体机构都推出检测技术，如何保障技术水平的统一性、行业认可的标准化等问题都指向媒体行业需要研发出行规性的检测技术。

其次，建立由主管部门牵头的AI研究小组或办公室。人机共生是未来媒体开展生产活动无法避开的话题。一是要从AI工具论走向研究论。媒体应该清晰地认识到AI信息收集、整理、生成的功能与媒体构建舆论场的功能高度一致，因此，媒体不仅要将AI作为新闻生产工具，还要将其作为研究对象，不能只依靠第三方AI软件介入媒体行业的活动，还要将AI发展的本身作为报道对象，而非展现它对其他行业“工具性”的效率化提升。二是需要研发AI电子追溯水印。现阶段，AI的生成“黑箱”导致过程不清，如果AI介入新闻生产流程，要如何把控与追责？根据新闻媒体的“留痕”机制可推导出，AI介入后如何“留痕”是当前应迫切研究的技术。现阶段，很多媒体对于AI作品都提供“AI生成”标识，但该标识是在使用AI希望被发现的“善意”主观性下才成立的，如果是在使用AI不想被发现的“恶意”主观性下就无法成立。这种情况下，数字追溯水印发挥的作用至关重要。这类技术含量高、财政投入大的研究，单一新闻机构难以做到，需要的是顶层设计。

完备的行规法规亟待落地。2025年9月1日，《人工智能生成合成内容标识办法》正式施行，要求网络信息服务提供者开展人工智能生成合成内容标识活动。但违反规定如何惩戒，文件规定较为模糊，当前缺乏关于利用AI所产生的各类纠纷该如何定量、定性的行政或法律规定。因此，媒体机构等信息服务平台要在保障数据安全前提下为政策制定提供样本，积极配合政府、主管部门定期提交脱敏后的AI介入记录，包括数据标注点位、内部审核日志，AI生成内容的原始材料和最终版本，为立法工作提供有效的实证性证据；同时要建立自己的数据库，对原始的稿件数据、修改的痕迹等信息进行归纳收集，由专门人员进行统一管理，以便后续相关责任的划分。

（作者汪茂盛系《当代党员》杂志社编辑部副主编，张侃琦系该杂志社总编辑）

责任编辑：喻　瑾

注释：

①余惠敏：《31亿用户构建全球最大AI试验场》，经济日报2025年8月3日。

②熊炎：《解释警示逆火效应是醍醐灌顶还是火上浇油？》，《新闻与传播研究》2019年第26期。