AI大模型遭恶意“投毒” GEO滥用侵蚀智能体公信力

▶ 本报记者 孙庆阳

GEO（生成式引擎优化），原本是帮助人工智能（AI）大模型准确找到真实信息的技术，但被滥用后就变成针对AI的恶意“投毒”。今年央视“3·15”晚会曝光了这一乱象：有人批量制造伪测评、虚构证据链，把营销软文伪装成客观知识，诱导大模型输出错误答案。这种行为会直接误导用户消费与决策，污染AI信息环境，严重破坏公众对人工智能的信任。

“投毒”背后的大模型漏洞

“从技术原理看，这次事件暴露的并不是大模型‘偶尔答错’。”中国信通院云大所人工智能卓越中心负责人连云波认为，当前不少大模型在外部信息接入链路方面缺少“真实性优先”处理架构。

北京交通大学法学院副教授、数据法学研究中心主任付新华揭示其逻辑陷阱：“当同一套虚假信息被多个账号、多个页面、多个模板重复发布后，模型可能会将这种‘人为制造的一致性’说法视为“多方印证”，但这些页面却可能来自同一套营销底稿和同一分发系统，只是换了站点、账号和标题；如果系统缺少来源分析、同源内容聚类去重和可信度分层，就会把“重复传播”误判为“独立证据”。

“最终，所有问题在答案生成阶段被彻底放大。”付新华表示，这与传统搜索有着本质区别。过去，用户能看到多个信息链接，可以自己分辨和判断；但生成式回答，如同一个善于整合的“故事大王”，会把那些零散、可疑的信息片段，通过强大的语言能力压缩、润色成一段流畅、肯定甚至带有鲜明建议色彩的完整“答案”。

例如，央视“3·15”晚会曝光的“Apollo-9”案例显示，仅靠批量生成和分发虚假软文，部分模型就会将虚构产品及其夸张卖点吸收进推荐答案。

合规与恶意如何界定

合规GEO与“投毒”的界限究竟在哪里？付新华判断：“凡是通过杜撰事实、伪造评测、冒充第三方立场、批量操控分发、刻意制造‘独立来源’假象来影响AI输出的行为，都应认定为恶意‘投毒’”。合规GEO，本质上是帮助AI更准确理解真实信息，如优化官网的表达，让产品参数、白皮书、权威报告更容易被AI发现和引用；而恶意“投毒”则是在人为制造虚假证据操控AI答案。

在实操层面，连云波为企业划出4条底线：一是内容必须真实可核验，不能虚构、不能夸大、不能把广告包装成评测；二是内容身份必须清晰，广告、生成内容、官方资料、用户评价不能故意混同；三是内容必须可追溯，能够说明作者、来源、发布时间、修改记录和支撑材料；四是在医疗、金融、教育、消费安全等高风险领域，应优先依赖备案信息、权威机构数据、正式说明书和资质文件，而不是依靠“内容矩阵”堆出虚假口碑。

反之，如果GEO滥用得不到遏制，付新华描绘了一种可能：“如果互联网上越来越多内容不是为了被人阅读，而是为了操控AI，那么真实内容、专业内容、慢生产内容就会在‘机器可操纵性’上输给批量制造的伪内容。”

“更可怕的是‘二次污染’。”连云波解释说，“今天污染的是搜索结果、检索语料和知识库，明天就可能通过再训练、蒸馏、模型更新形成‘二次污染’。”

“最深层的损害或许是信任。”付新华认为，一旦公众普遍形成AI推荐也不过是高级软文广告的印象，那么大模型就很难再承担起知识助手、消费顾问、决策参考等更高层次的功能。

GEO滥用的应对路径

“对于平台而言，关键在于建立信息免疫系统。”付新华认为，前端是“信源治理”，对官网、监管信息、主流媒体、自媒体进行差异化的可信度分层，让权威声音拥有更高权重。中端是“异常识别”，系统需要能识别出那些短期内集中发布、表述高度雷同、相互循环引用、带有明显商业操控意图的“内容集群”，防止模型被批量伪造的一致性所欺骗。后端则是“响应处置”，一旦出现重大风险事件，相关平台要有能力迅速启动“下架—降权—纠偏—回滚”的处置机制，避免错误答案长时间流传。

连云波建议，真正可持续的路径，是建设高质量官网、白皮书、产品文档、服务记录和知识库，让AI“有根据地引用”。只有平台提升治理能力、企业守住真实底线、用户保持基本核验习惯，AI信息生态才能逐步建立长期信任。

而对于普通用户，付新华建议把AI当作“线索工具”，而非“最终裁判”。尤其在涉及消费决策时，要多做几步核验：一要追问来源，不要只看AI给出的结论，要继续追问“这个推荐依据什么信息”“有没有官方来源或独立测评”。二要做交叉验证，把AI推荐和品牌官网、电商平台详情页、主流媒体报道、第三方专业测评对照起来看，不能只凭一轮问答就下单。三要警惕极端表述，凡是“第一名”“唯一推荐”“全网口碑最好”“闭眼买”这类绝对化措辞，都要提高警觉。四要优先看可核验事实，如参数、认证、售后、价格区间、用户真实评价，而不是被包装出来的故事性话术。

海淀北部打造人工智能产业重要策源地

▶ 本报记者 邓淑华

作为全国科技创新核心区，北京市海淀区汇集人工智能（AI）企业超过1900家，AI核心产业规模近3600亿元，北京市近70%的AI大模型诞生于此。

记者近日获悉，作为北京市海淀区强化战略腹地和高精尖产业承载关键区，海淀北部片区（以下简称“海淀北部”）汇聚了从AI基础研究到AI场景应用的全链条资源，集聚了芯片设计、算法框架、算力基础设施、行业大模型等关键环节的头部企业，这里正在成为北京市乃至全国AI发展的重要策源地和产业集聚区。

“北部片区立足自身优势，紧扣AI产业发展布局，聚焦算法、算力、数据三大核心要素，培育新赛道，做强产业集群。”北京市海淀区副区长、中关村科学城管委会副主任唐超介绍说。

助力解决千行百业痛点

柔性精密制造是高端装备制造的基石，但普遍面临小批量、多品种、定制化等生产困境。该行业企业常需耗费数月开发生产程序，却仅获得个位数订单，投入与产出严重失衡，容易陷入招工难、建厂难、扩产难的困境。

“针对这些痛点，我们通过具身智能，借助AI的自我学习和快速适应能力，重新定义柔性精密制造。”北京大学先进制造与机器人学院长聘教授、一湃（北京）智能科技有限公司合伙人庞智博表示。

在海淀北部，这样的探索还有很多。北京中科慧灵机器人技术有限公司通过井下机器人适配自动化设备以及智能开采平台，搭建了一整套井下无人化作业架构；深腔（北京）科技有限公司研发的管道机器人，专为燃气、热力等管网进行内部检测设计，并使之适应复杂管道环境；金扇叶（北京）健康科技有限公司研发出中医经络监测仪、舌面分析系统、体质辨识仪以及医用红外热像仪，能够早期发现潜在健康问题……

AI的价值正从宽泛走向专精，从对话走向实干，助力解决千行百业的痛点。

聚焦三大核心要素

据悉，海淀北部正在以算法突破筑牢技术根基、以普惠算力夯实产业底座、以高质量数据激活应用场景，推动AI与实体经济深度融合。

海淀北部片区供图

中科慧灵生产的人形机器人，在北京市海淀北部片区人工智能创新生态发展论坛上做展示。

“当前，航空航天航海等领域高度依赖研发设计类工业软件，然而，CAD（计算机辅助设计）、CAE（计算机辅助工程）等核心工具长期被国外垄断，我们希望打破这样的瓶颈。”北京大学力学与工程科学学院院长杨越介绍说，“北京大学依托理工科优势，进一步聚焦工业软件前沿基础理论算法，引领新一代颠覆性技术，形成自主可控的软件。我们重点布局四大方向：一是研发设计类软件的共性根技术；二是开发更面向前沿的云边端协同智能工业软件；三是构建量质融合的CAE求解新范式；四是对典型行业进行赋能，其中最需要硬核攻关的是‘三深三极’（深空、深海、深地；极端环境、极端过程、极端尺寸）极端场景专用工业软件，相关成果曾获戈登·贝尔奖。”

算法是AI的“大脑”，而算力正在向专用算力延伸。

“我们的产品覆盖具身智能、生物医药、教育科研、生态工艺制造等工业领域。”北京九章云极科技股份有限公司相关负责人王凯凯表示，该企业通过灵活的算力调度与计费模式，降低AI应用门槛，让算力像水电一样普惠千行百业的具体场景。

高价值数据是特定领域的高质量交互数据。北京中科慧灵机器人技术有限公司董事长张正涛表示，相较大语言模型所需的海量互联网文本数据，具身智能需要的数据集是面向真实开放场景、多模态的物理交互数据。“我们提出了基于MR（混合现实）沉浸式的数据采集，打通了从仿真数据到真机数据的采集模式，管理大量高质量数据。”

“‘十五五’开局，北京市海淀区委提出‘五方六力’成果转化机制。”唐超提出，希望海淀北部锚定核心战略，打造产业标杆；深化服务下沉，做“五方六力”的协同者；携手同行，争做启智新程的同路人。