“目前正在积极调查当中。”3月16日,北京市怀柔区市场监督管理局工作人员的一则简短回应,拉开了针对“AI投毒”黑产监管整顿的序幕。

就在前一晚,央视3·15晚会曝光了利用GEO技术给AI大模型“投毒”的黑色产业链,北京力思文化传媒有限公司运营的“力擎GEO优化系统”被点名。这套系统通过批量生成虚假宣传软文并自动发布至多个自媒体平台,在短短几天内就能让一款完全虚构的产品被主流AI助手当作真实信息推荐给用户。

而在无形的互联网端,一场关于“AI竟然还能投毒”的认知风暴,则早在报道播出的一刻已在疯狂吹袭。

■新快报记者 郑志辉

3·15曝光“AI投毒”后追踪

深度解析“认知操纵”黑灰产业链

“Apollo-9”还在说话

晚会刚曝光完,有网友扭头就去问了几个AI大模型同一个问题:“Apollo-9智能手环怎么样?”

Apollo-9,正是记者在晚会演示中虚构的那款产品,带有“量子纠缠传感”“黑洞级续航”等夸张描述,本应只存在于“力擎”系统批量生成的软文里。

测试结果令人惊讶:有的AI模型反应迅速,给出了清晰答案:“这是虚假宣传的典型案例,已被2026年3·15晚会曝光。”有的模型虽然没有直接提及晚会,但通过逻辑判断得出结论:这款产品不在任何主流品牌产品线中,可能是小众品牌或山寨产品。

然而,还有AI模型仍在“中毒”——它们一本正经地向用户介绍着这款根本不存在的智能手环,复述着那些杜撰的“用户好评”。

网经社资深人工智能专家郭涛向记者解释了此次曝光案例的特殊性:“传统的AI投毒多是在训练数据中植入恶意数据,如定向后门投毒或非定向污染投毒。而此次案例是利用GEO技术,通过在互联网上系统性、定向投放大量虚假信息,让AI大模型将其捕获并作为答案输出,本质是对AI信息生态的污染,是一种新的认知操纵方式。”

被曝光的“力擎GEO优化系统”运营者李总在暗访中直言:“GEO业务受热捧的主要原因就是它能在AI大模型里帮客户‘喂料投毒’,实现客户的商业目的。”

这又带出了业内竟然还有部分服务商在提供“抹黑竞品”(黑公关)服务,即通过向AI投喂虚假或者污蔑信息,来干扰竞争对手的搜索表现。

如此一来,当前全球热烈追捧的生成式AI,竟然沦为黑产的“广告机”甚至“喷粪机”。给AI投毒,也就形成了围猎AI、获取暴利的新的黑灰产业链。

谁在给AI“下毒”?

3·15晚会的相关报道,另外一件震撼普通消费者的事情是,短短几天时间,一篇虚构产品的软文就能让AI信以为真并“倾力推荐”,这是什么效率?这还能叫AI吗?

蚂蚁集团大模型安全专家毛宏亮坦言,只要大模型依赖外部数据学习,就可能被“喂毒”。这次的曝光,只是让这种潜伏的风险彻底摆到了台面上。

而这次“投毒”来得既快又猛,他认为,主要有三个推手:首先是“造假”变得太便宜了,现在有了生成式AI,制造海量“毒性内容”的成本几乎为零,晚会上提到的那些被检索到的网页,一眼就能看出是AI批量生成的“工业流水线产品”;

其次,搜索引擎成了“帮凶”。现在的AI助手都连着网,能实时搜索。黑客只要利用搜索引擎的规则,让那些假网页在搜索结果里排名靠前(高热召回),AI在回答问题时就会自动抓取并引用这些假信息。

此外,黑客现在不仅造假内容,还研究出了一套专门写给AI看的“八股文”。他们在假文章里刻意加上“权威来源”“结构化分段”等特征,专门迎合大模型的阅读喜好。这就像给AI准备了它最爱吃的“特制毒药”,让它更容易误以为是真知灼见而全盘吸收。

天眼查App显示,“力擎GEO优化系统”关联公司北京力思文化传媒有限公司成立于2018年4月,法定代表人李千钟全资持股。今年2月,该公司登记“媒体发文及管理平台”软件著作权,近期还申请注册了“壹豹客”“力擎”商标,国际分类涉及网站服务、科学仪器。

GEO与“投毒”之辨

随着“AI投毒”事件持续发酵,有必要厘清一下GEO优化与“AI投毒”行为的区别。

GEO,全称生成式引擎优化(Generative Engine Optimization),是一套专门针对AI平台的内容优化策略。它的核心目标是提升品牌信息在AI生成答案中被引用和推荐的几率,使企业内容被AI算法识别为“可信来源”。通俗理解就是,让内容更容易被AI看懂、采信的一种优化方式。

云南辛树计算机技术有限公司总经理、科普中国专家李羿鹏指出,“GEO优化本身没有善恶,决定它好坏的,是使用它的人。”

这让当前GEO行业出现了所谓的“白帽GEO”和“黑帽GEO”之分。前者遵循规则、基于真实信息进行优化;后者则通过批量生成低质甚至虚假内容、编造排行榜等方式污染大模型信息源,干扰AI输出结果,也就是此次3·15晚会曝光的“AI投毒”。

3月16日,多家GEO行业相关企业密集发布声明。

AB客发布声明称:“坚决反对任何形式的虚假宣传、数据造假、恶意操控AI输出结果的行为,不参与、不开发、不提供所谓‘洗脑AI’‘操控标准答案’‘一周见效刷排名’等违规服务。”

作为科大讯飞生态伙伴的河南恒辉合焕网络科技有限公司也紧急发声:“坚决杜绝虚假信息、批量软文造假、AI数据投毒,不使用黑帽手段、刷量刷排名、恶意劫持、诋毁竞品,不做‘保底首页、永久霸屏、绝对效果’等虚假承诺。”

防御战打响

面对这种新型黑灰产业链,业界是否有能力防御?

郭涛向记者介绍了当前的防范思路:“业界主要从以下方面防范:开发者与服务提供者严格管控训练数据来源,建立数据清洗与脱敏机制,在训练过程中部署异常检测机制,上线前进行安全对抗测试,建立常态化的模型运营监控。”

他特别提到开源社区的作用:“开源社区要构建开源生态安全屏障,建立数据集、模型权重的安全审核与备案机制,普及AI安全知识。”

但郭涛也坦言,目前仍存在难以完全防范的漏洞:“互联网信息海量且繁杂,难以对所有信息进行实时、全面的真实性校验和监控。一些恶意信息可能隐藏在正常信息之中,不易被轻易识别。此外,AI模型的算法不断更新,攻击者也会随之不断改进投毒手段,防御措施可能无法及时跟上攻击技术的变化。”

毛宏亮介绍说,技术层面,可以通过架设三道“防火墙”来进行防范:“查户口”(信源分层)、“找破绽”(数据体检)和“请外援”(引入权威辟谣平台作事实核查)。

但当前业界面临的最大痛点是攻防成本严重“倒挂”:黑客动动手指,用AI几分钟就能生成成千上万条假新闻,成本极低,花样翻新极快;而防守方(平台)需要建立庞大的审核系统、购买权威数据、训练鉴别模型,每一分钱的治理成本都远高于对方的攻击成本,“这是一场不对称的战争。”他说。

普通人的“防毒”指南

李羿鹏表示,对消费者来说,我们问AI,图的是方便、客观、真实。可一旦推荐被投毒,你看到的就不是建议,而是包装得很像真话的广告。假产品、假功效、假口碑、假测评,普通人很难分辨。轻则花冤枉钱,重则买到不安全、不合规的商品,等到想维权时,往往找不到源头,投诉无门。长期下去,被破坏的不只是信息环境,还有大家对AI、对数字内容的整体信任。

面对这场AI信任危机,普通用户该怎么办?

郭涛给出了一套实用的“鉴别指南”:“用户可通过以下方法识别:追问AI答案的来源,若其含糊其辞则需警惕;对于‘推荐类’问题要提高警惕,如包含‘推荐’‘哪个好’等字眼的问题答案;搜索AI推荐产品的名字,查看是否有大量雷同软文。”

他特别提醒:“注意AI回答中的‘确定性语气’,若全是正面且无缺点的‘完美推荐’需小心;用‘反向提问’测试AI立场;去电商平台评论、垂直社区等‘非AI来源’交叉验证;对不知名产品的AI推荐保持最高戒备。”

湖北赋兮律师事务所但丁律师则从维权角度给出建议:如果消费者掉进AI“种草”陷阱,关键在于保存证据——对AI对话进行录屏,完整录下从提问到获得虚假推荐的全过程;对商品页面截图,保存订单详情和商家信息。然后再向平台投诉或依据消费者权益保护法维权。