责任编辑:张 桐 今晚报 21
副刊专刊部电话:23602884 2025年5月20日 星期二

利用AI造谣获罪,记者在看守所对话犯罪嫌疑人王某某——

30余万条不实帖文都由AI生成

“我不知道上海有一家华山医院,也不知道‘张明远’是谁,(内容)都是AI生成的。”

“我们有500多个账号,每天发1万多条帖文。”

“直到警察找来,我都不知道是哪篇帖子出问题了。”

今年3月,一则内容为“华山医院前院长张明远因阑尾炎未得到及时救治,客死他乡医院走廊长椅”的图文信息,一度在社交平台和微信群刷屏,造成不良社会影响。

上海公安部门发现这一情况后,立即展开调查:静安公安分局到华山医院询问情况,核实后发现该院不仅没有名叫“张明远”的前院长,甚至没有叫这个名字的职工;网安总队利用技术手段开始追溯,消息来源指向某网络信息平台的自媒体账号。最终结果显示,该网帖内容为不实信息,涉嫌造谣。

掌握相关证据后,静安警方将犯罪嫌疑人王某某、郭某某和石某某三人抓获。目前,三人因涉嫌非法利用信息网络罪,被上海警方依法采取刑事强制措施,相关违法账号被依法封禁。

记者在看守所对话了王某某。据了解,由王某某创立的三人工作室,共控制了500多个自媒体账号,日均发帖万余条,累计发布不实帖文30余万条。而这些帖文内容,均由AI改写或生成。他们摸索着平台的流量密码,设置关键词让AI产生爆款内容,以此获取分成收益。至于内容的真实性,似乎不在他们的考虑范围内。

谣言内容都用AI编的

记者:这条谣言是怎么产生的?

王某某:(内容)都是由AI生成的。我不知道上海有一家华山医院,也不知道“张明远”是谁。具体是什么内容,到现在我也不清楚,我们是批量采集和发布的。

记者:生成这条谣言时,给AI下达了哪些指令?

王某某:有一个App,是我们的素材库,可以按照不同领域、阅读量、发布时间等维度搜索平台的爆款帖文。人工采集这些内容后,我们就让AI软件进行批量改写或生成新内容,之后再用软件发到平台上。

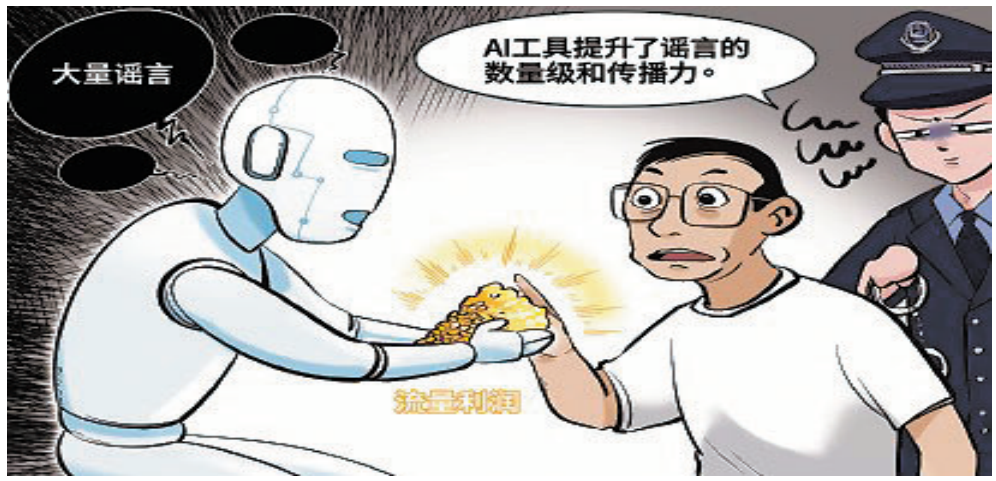
当时,让AI生成这条信息的具体指令,我不记得了。一般来说,我们会输入爆款帖文的原文,然后要求AI变换语序、语气以及关键词,生成新的文字内容。另一种,就是用AI的联网功能,让它根据爆款帖文的中心思想,先在网上搜索关联信息,再生成新的文字内容。

记者:你认为自己发布的内容属于什么性质的?

王某某:算是综合的,有些是新闻,有些是观点,主要还是按热度榜的话题和内容来分类。

记者:对于AI生成的内容,会审核真伪吗?

王某某:当时真的没有考虑那么多,我想的是,我们发的内容都是别人发过的,只不过做了加工,而且我们也是从平台热度榜单采集的素材,感觉是比较权威的。站在我们的角度,需要审核的内容,就是根据平台的要求,删除一些敏感词。说实话,直到警察找上门来,我都不知道是哪篇帖子出了问题。



日均编造上万条帖文

记者:你们会标注内容是由AI生成的吗?

王某某:不会。平台也没有这方面的要求。只要没有敏感词、涉黄涉暴画面等,就能通过平台审核。我没有遇到过平台对文字内容出现违规提示。据我了解,(自媒体)这个圈子里,绝大多数人都是这样做的。

记者:如何靠发文获利?

王某某:我们的账号每条帖文在200—500字。我们有4个“大号”,还有500多个“小号”。“大号”是跟MCN合作的机构账号,每个“大号”可以添加“小号”作为子账号,子账号获得的收益再流转到母账号,根据流量情况,每个月从平台分成获益。

记者:收入跟流量的关系是什么?

王某某:对我们来说,流量就意味着收入。我们每天发布1万多条帖文,内容要有些差异,AI都能完成。然后再用API端口自动发布到不同账号里。具体流量说不准,有的帖文有几万阅读量,最高的一条能有十几万。阅读量达到1万,我们就能有5元左右的收益。从2月底开始到被抓,我们总共有4万多元的收益,但还要跟MCN机构分成。收入并不高。

编造明星八卦被警示

记者:哪些内容流量比较大?

王某某:八卦类的内容,比如明星八卦、社会八卦,还有涉及职场、科技、养老、国际、教育、医疗等这些领域内容,都比较吸睛。现在想来,像“华山医院前院长”这条,可能重叠涵盖了社会、职场、医疗、国际等领域,所以才得到关注。

记者:之前用AI生成的内容,有被判定为谣言的吗?

王某某:某女明星刚去世时,对于孩子抚养权的问题,我们也用AI生成过内容。后来被平台判定为谣言,有过警示,也被扣分处罚了。像“已过时效”“标题党”这些,也会被平台处罚,但我们还是有侥幸心理。其实,相比我们生成发布的内容数量,被平台判别处罚的内容比例非常低。

记者:你认为你们发布的哪些内容算谣言?

王某某:以前没想过这个问题。这几天在反思,我们用AI生成了很多没有事实根据的、带来不良社会影响的内容。

追逐流量要有底线

记者:你是怎么进入自媒体行业的?

王某某:我以前是做电商的,自己没有货源,相当于做别人的代理。但这两年效益不好,利润低、成本高,然后就想着转换赛道。后来,互联网圈里的朋友推荐我做自媒体。使用的软件和工具在网上都能找到,方法也有人教,只要花时间进去,就能赚钱,成本很低。

记者:你认为追逐流量需要底线吗?

王某某:肯定要有底线,不应该胡编乱造或者夸大其词,不应该编造发布虚假信息。我现在认识到了以前做得不足的地方,没有把握好尺度,也没有好好审核。

记者:对于自己做的事情,有什么反思?

王某某:经过警方的教育,我知道自己错在哪里,也感到非常后悔。我会接受应有的惩罚,吸取教训。

“AI造谣”带来哪些影响?如何阻击?

近年来,AI造谣在全国多地频发,传播快、门槛低、迷惑性强的特征日益显现。比如,在一起女童走失事件中,一团伙以“标题党”“震惊体”方式,恶意编造谣言。经查,该团伙利用AI工具等生成谣言内容,通过114个账号矩阵,在6天内发布268篇文章,多篇文章点击量超过100万次。

如今,不断升级的AI工具,让谣言产生的数量级和传播力几何级增长。“AI的介入,使得信息的生成和传播速度大大加快,同时也增加了判断信息真实性的难度。”结合“华山医院前院长客死异国”这一谣言的查处情况,静安公安分局网安支队民警程海岳表示,当下,AI谣言已能够定制化生成、精准化传播和智能化扩散。

跟王某某等人的出发点一样,近来发生的多起AI谣言,造谣者的动机是为了发布帖文、博取流量,以获取互联网内容平台给予创作者的点击量、阅读量奖励。此外,也有一些人通过AI造谣来“养号”,为直播和卖货引流。上海公安机关曾发现,在一家电商平台上出现了某艺人“命运多舛、含恨离世”等短视频,引发大量点赞和转发。经查,该视频内容完全系伪造。视频发布者到案后交代,他在某电商平台上经营一家土特产网店。由于销量不佳,他便通过编造夺人眼球的虚假新闻给网店账号吸引流量。

应对AI造谣乱象,除了强化打击和处罚力度外,强有力的制度规制和综合治理同样重要。今年3月,《生成式人工智能服务管理暂行办法》发布,明确“生成式AI服务不得生成法律禁止的内容”,包括虚假信息等;若企业未履行内容审核义务,导致谣言传播,将面临最高500万元罚款,直接责任人处10万元以下罚款。去年4月发布的《关于开展“清朗·整治‘自媒体’无底线博流量”专项行动的通知》,已要求加强信息来源标注展示;使用AI等技术生成信息的,必须明确标注系技术生成;发布含有虚构、演绎等内容的,必须明确加注虚构标签。

在监管层面,对网络谣言的发现和查处手段也在升级优化,以适应AI造谣的新挑战。上海公安部门不断强化“人工+技术”手段,加强与互联网内容平台的联动共治,并利用大数据和应用模型,全量清理造谣传谣信息,分级惩治涉谣账号,严厉打击“造热点”“带节奏”的网络水军。今年以来,已破获网络水军案件15起,抓获38名违法犯罪嫌疑人。

邬林桦 摘自《解放日报》