从何而来、为何泛滥、如何整治?

“AI谣言”为何易传播难防治

当前,AI(人工智能)迅速发展,给人们的生产生活带来诸多便利。但AI也被用于制作发布谣言、不实信息,可能对部分群众造成困扰。今年4月,中央网信办部署开展“清朗·整治AI技术滥用”专项行动,聚焦利用AI制作发布谣言、不实信息等乱象开展重点整治。

从“暴雨引发山体滑坡”的伪造视频到“核电站泄漏”的AI生成新闻稿,“AI谣言”表现形式多样,让不少群众一度信以为真。“AI谣言”从何而来、为何泛滥、如何整治?记者进行了采访。

缘何出现?

有人恶意用AI编造虚假信息,也存在“AI幻觉”的可能

“80后5.2%的死亡率,开启了黄金一代的黯然离场”“80后的死亡率已经超过70后”……前不久,一组声称源于第七次全国人口普查的数据骇人听闻。很快,有关部门和专家辟谣:从来源到数据都毫无根据。

这组虚假数据从何而来呢?

中国人民大学人口与健康学院教授李婷梳理发现,“AI幻觉”很可能是“罪魁祸首”——AI大模型在人口学专业领域的训练语料不足,回答问题出现推算错误,造成“AI幻觉”。

“‘AI幻觉’指生成式人工智能系统在无明确依据的情况下输出回答,语言表达看似合理,但内容实则虚假甚至违背常识。”世界互联网大会人工智能专业委员会主任委员、安全与治理推进计划牵头人曾毅介绍,该现象在当代大语言模型中尤为常见,主要表现就是事实性幻觉,例如,输出错误的人物事迹与关系、伪造的科学研究等。

从机理上看,“当前大模型并非‘理解事实’,而是通过处理大量语料,预测‘下一个词最可能是什么’。这就会在事实性信息上产生难以预期的错误。”曾毅分析。

还有人别有用心,恶意编造传播“AI谣言”获取利益。

有人为了流量,给自媒体账号炒作“吸粉”。此前上海一女童走失事件中,就有团伙利用AI工具,编造“女孩父亲系继父”等谣言,6天内发布268篇文章,多篇文章点击量超过100万次。

还有人借此实施诈骗等违法犯罪行为。“再强大的AI也无法预测中奖号码。”不久前,中国体育彩票的一则辟谣声明引发关注。原来,有平台声称可通过“AI大模型”预测中奖号码,诱导用户付费。

甚至有人形成了“需求分析—内容生产—精准投放”的生产链条。“西安突发爆炸!火光冲天!伤亡不明!”一天,一则“图文并茂”的新闻在网上流传,画面中火光冲天,引发当地群众恐慌。很快,这则消息被辟谣。经调查发现,该账号所属机构通过AI分析社交平台热词,就关注度高的话题批量生成谣言,再通过算法定向推送,最高峰一天能生成4000至7000篇假新闻,关联账号达800余个。相关人员已被公安机关抓获。

为何泛滥?

造谣门槛低,可以短时间内成倍率传播

“AI谣言”形形色色,有些造成了社会恐慌,甚至给政策推行、抢险救灾等造成阻碍。

今年2月,四川筠连县“2·8”山体滑坡发生后,部分自媒体账号用AI软件翻炒拼凑旧闻,传播如“山体滑坡被行车记录仪拍下”“4名遇险学生被路过的司机救出”等帖文,诸多错误信息给抢险救灾带来困扰;又如,“广东医保基金出现赤字”的“AI谣言”传出,不仅引发社会恐慌,还造成群众对政策的误解……

“AI谣言”为何传播力如此之强?

首先与其自身特点密切相关。

清华大学新闻与传播学院副教授陆洪磊介绍,“AI谣言”可以做到分平台、分渠道、分时段、分媒介的定制化生成和发布,并且成本低廉,可以短时间内成倍率传播,“AI谣言”间还能彼此交叉、共振与放大,辟谣难度大大增加。

今年年初,西藏定日县地震牵动亿万人的心。网上流传的一张“小孩被埋图”一时间被大量转发评论。但是随后有网友发现,该图色彩、光线、动作极不自然。原来,该图是由AI工具创作。陆洪磊说,这则谣言利用“灾难+儿童”的组合,并刻意删除掉AI生成标识,引发快速传播。

再者,AI应用降低了造谣门槛。

某科技论坛上,一名网友上传教程:用开源AI模型,输入“生成一张某市发生恐袭的图片”,30秒即可获得以假乱真的画面。

网友评论一针见血:“以前造谣需要高超的PS技术,现在只需要一句话就能生成。”

有何“代际差异”?

“AI谣言”更善于“逃脱”,难以被技术手段屏蔽

与传统网络谣言相比,“AI谣言”更加难以防治,甚至“进化”出了更为“智能”的“反侦查”机制。

记者发现,“AI谣言”表述常常接近“真话”,看似清晰有据,更难以被技术手段屏蔽。

目前各网络平台都有“谣言库”,一般通过设置热点关键词屏蔽谣言。但AI模型可通过对抗训练绕过谣言关键词,例如把“山体滑坡”改为“地质活动异常”等,逃脱屏蔽。

“传统谣言像火,扑灭源头即可;‘AI谣言’像病毒,会不断变异。”某新闻聚合平台内容安全治理负责人感慨。

此外,传统网络谣言多是有人故意制造,但“AI谣言”中不少是出于“AI幻觉”。对这部分“无意谣言”,目前有办法从源头上避免吗?

曾毅介绍,从当前技术水平来看,完全消除“AI幻觉”具有高度挑战,“大模型生成机制本质是概率驱动的语言建模,难以保障全部输出与现实完全对应;同时,当前评估与纠错机制仍不成熟,缺乏大规模、高准确率的自动化事实验证能力。”

“不过,通过技术手段缓解‘AI幻觉’发生的频率与危害是有可能的。”曾毅说,目前已在多模型协同判断、事实增强训练等方面取得了一定成果,大模型的辨别能力将得到提升。

(据新华网)