

当AI“一本正经胡说八道”……

新华社记者 颜之宏 胡林果

当前, AI正赋能千行百业, 为人们的工作、学习、生活带来极大便利。与此同时, 不少人发现, 用AI搜索数据, 给出的内容查无实据; 用AI辅助诊疗, 出现误判干扰正常治疗……AI频频上演“一本正经胡说八道”。社交平台上, AI幻觉引发热议。

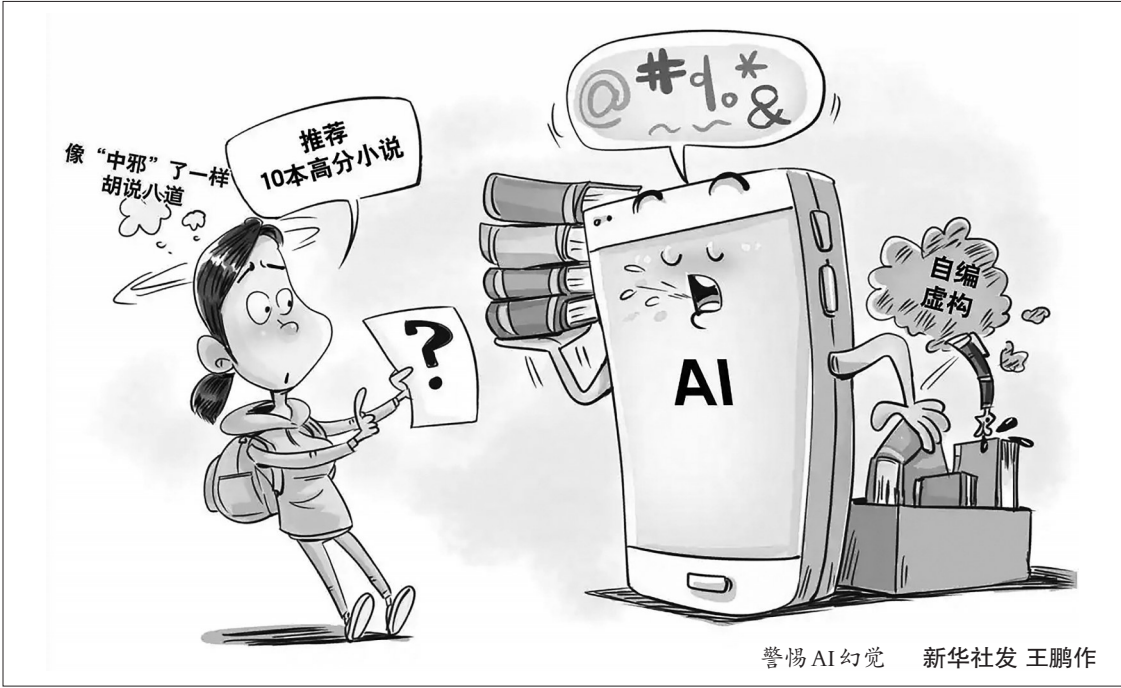
AI好用但不时像是“中邪”了

用AI检索海量信息、让AI辅助查看三维病灶、打造AI互动课堂……如今, AI已深度融入现代生活, “人工智能+”产品赋能各行各业, 从多个维度提供便利。作为AI深度使用者, “95后”女生瑞希坦言, AI好用, 但不时像“中邪”了一样胡说八道。“我让AI推荐10本高分小说, 结果一多半都是它编的。反复确认后, 它承认虚构了答案。”

现实生活中, 不少人遇到相似情况。业内人士表示, 这是AI幻觉导致。“AI可以快速给出答案, 但生成内容可能与可验证事实不符, 即凭空捏造; 或生成内容与上下文缺乏关联, 即‘答非所问’。”一名主流人工智能厂商技术人员说。记者使用一款AI软件, 让其给出某行业未来市场规模及信源, AI迅速回答称某投资机构预测2028年该行业的市场规模将达到5万亿美元, 并提供相关链接, 但链接页面找不到上述信息。记者看到, 页面内容虽然包含该投资机构名称和5万亿美元表述, 但预测数据并非该机构作出, 且不存在2028年时间节点。

社交平台上, AI幻觉相关话题浏览量达数百万, 网友吐槽涉及金融、法律、医疗、学术等多个领域。

第三方咨询公司麦可思研究院近期发布的2025年高校师生AI应用及素养研究显示, 四千余名受访高校师生中, 近八成遇到过AI



幻觉。今年2月, 清华大学新媒沈阳团队发布的报告指出, 市场上多个热门大模型在事实性幻觉评测中幻觉率超过19%。

AI幻觉已经影响了人们的生活与工作。

近期, 一名国外男子被诊断出溴中毒。他此前询问AI, 过量食用食盐不利于身体健康, 有无食盐替代品, AI回答称可以用溴化钠代替。但溴化钠存在一定毒性, 需要严格遵医嘱服用。该男子用溴化钠代替食盐三个月后出现精神错乱等症状。

这几年, 美国多起案件中的律师因在法律文件中使用AI生成的虚假信息, 被法院警告或处分。

AI幻觉为什么会发生?

受访专家认为, AI幻觉的背后存在多重因素。

——数据污染。AI“养成”过程中, 数据“投喂”是关键环节。研究显示, 当训练数据中仅有0.01%的虚

假文本时, 模型输出的有害内容会增加11.2%; 即使是0.001%的虚假文本, 其有害输出也会相应上升7.2%。

奇安信集团行业安全研究中心主任裴智勇解释说, 人工智能大模型需要海量数据, 训练数据来自开源网络, 难免会错误学习一些虚假、谬误数据, 还有一些不法分子会恶意进行“数据投毒”。

“如果把AI比作一个学生, 数据污染就像是给学生看了错误的教科书, 自然会导致‘胡说八道’。”暨南大学网络空间安全学院教授翁健说。

——AI本身“认知边界模糊”。翁健认为, 人类智能的一个重要特征是“元认知”能力——知道自己懂什么、不懂什么, 而当前AI技术架构缺乏这种自我认知机制。

翁健解释称, AI可以博览群书, 但并不一定理解书里的内容, 只是根据统计规律把最有可能的

词语组合在一起, 在准确评估自身输出的可信度方面尚存盲点。

——人为调校和干预。在中国通信学会数据安全专业委员会副主任委员左晓栋看来, 相较于事实真相, AI更在意自己的回答是否契合用户需求, 从而导致AI有时为了“讨好”用户而编造答案。

“针对不同需求, AI的训练、打分方式也不同。”一位从事大模型训练的技术人员说, 当面对写作等创意性需求时, 偏理性的事实严谨在打分系统中占比相对较低, 偏感性的词语优美、富有感情色彩等占比更高。“所以可能会出现一篇辞藻华丽但词不达意的文章, 里面内容甚至与事实相悖。”

多方合力减少AI幻觉

第55次《中国互联网络发展状况统计报告》显示, 截至去年12月, 有2.49亿人使用过生成式人工智能产品, 占整体人口的17.7%。受访专家表示, 应通过多方合力应

对AI幻觉带来的风险挑战。

今年4月, 中央网信办印发通知, 在全国范围内部署开展“清朗·整治AI技术滥用”专项行动, 训练语料管理不严、未落实内容标识要求、利用AI制作发布谣言等均为整治重点。

“可靠、可信、高质量的数据对降低AI幻觉非常重要, 应优化人工智能的训练语料, 用‘好数据’生成‘优质内容’。”左晓栋认为, 可以加快推动线下数据电子化, 增加“投喂”的数据量; 同时探索建立具有权威性的公共数据共享平台, “各大厂商也应加强优质数据筛选, 提升训练准确性”。

多家主流人工智能厂商已经采取措施, 从技术层面减少AI幻觉发生。

豆包升级深度思考功能, 由先搜后想变为边想边搜, 思考过程中可以基于推理多次调用工具、搜索信息, 回复质量明显提升; 通义千问在20多个通用任务上应用强化学习, 增强通用能力的同时纠正不良行为; 元宝持续扩充引入各领域的权威信源, 在回答时交叉校验相关信息, 提高生成内容的可靠性。

翁健建议, 建立国家级人工智能安全评测平台, 就像生物医药新药上市前要做临床试验一样, 大模型也应该经过严格测试; 同时, 相关平台加强AI生成内容审核, 提升检测鉴别仍能力。

“AI可能‘欺骗’用户, 公众应客观认识人工智能的局限性。”左晓栋等专家提示, 可以通过改进使用方式, 如给出更加明确的提示词、限定范围等避免AI幻觉。“无论是工作、学习还是生活, 现阶段的人工智能还不能全面替代人类的认知和创造能力, 大家在使用AI时, 要保持怀疑态度和批判思维, 不过度依赖AI给出的回答, 多渠道验证核查。”

(新华社广州9月24日电)

人工智能时代电影行业的变与不变

新华社记者 陈怡 孙一然

22日, 一场名为“未来影像”的AI电影国际峰会在第30届釜山国际电影节期间举办, 引发业内人士关于AI对电影行业影响的热议和深思。如大家所见, AI正以前所未有的速度和广度渗透电影工业, 带来一场堪比工业革命的深刻变革; 但不少电影人坚持, 电影创作的核心本质无法被取代。

釜山国际电影节官方产业交易平台“亚洲内容与电影市场”在今年推出全新单元InnoAsia, 将AI等前沿科技纳入视野, 探索AI如何重塑电影表达。此次AI电影国际峰会由字节跳动旗下AI创作与内容平台即梦AI、云和AI服务平台火山引擎和上海电影股份有限公司等机构共同举办, 此前他们联合发起的“未来影像计划”全球AI影像作品征集活动的最终获奖作品, 也被带到InnoAsia平台展映。

此次“未来影像计划”获奖短片的创作均使用即梦AI等生成式人工智能创作平台。依托字节跳动自研的Seedance视频生成模型, 即梦AI能够帮助创作者突破技术门槛, 实现“创意即产出”。火山引擎则为创作过程提供了底层技术支持, 其云端算力集群能够保障Seedance模型的高效运算。

即梦AI产品经理林捷表示, “国内有着非常繁荣的短视频生态, (在技术层) 会比较紧密地去向创作者收集各种各样的反馈, 创作者也会反过来帮助我们去快速迭代模型。”

不论是短视频还是大电影, AI技术正以“全链条渗透”方式融入影视产业生态。火山引擎旗下大模型服务平台火山方舟模型策略产品高级总监向凯接受新华社记者采访时表示, 目前AI在影视里的应用虽然还处于初级阶段, 但几乎覆盖了电影制作的全部流程。AI开拓了创作者们的自由度, 助其提升效率。

参加峰会的博纳影业集团影视制作副总经理曲吉小江告诉记者, “AI技术逐步成熟, 整个业界从初期的怀疑排斥到现在慢慢地接受、尊重并使用, AI生成影视可以说快迎来要爆发的‘奇点’”。

曲吉小江在当天的峰会演讲中说, AI正重构电影制片模式, 将原本难以把控的人物、造型、光线等变得“精准可控”, 符合电影工业化指标。她表示, AI不仅仅是一种工具, 更像是人类的“外脑”, 赋予了创作者“超能力”。与历次电影所面临的技术变革不一样, AI将彻底改变人们的思维方式。

上海温哥华电影学院电影制作系高级讲师奥黛·阿瓦迪亚也认为, 利用AI技术将是影视行业的主流趋势。“尽管好莱坞对于AI评价褒贬不一, 但不能否认的是, 越来越多的制片人会使用AI。AI能节省预算, 也能缩短工期, 有了AI, 可以尝试更多可能。”

然而, 在AI给影视行业带来的巨大变化中, 也有声音认为AI无法改变电影创作的核心。此次“未来影像计划”获奖短片导演——韩国影视创作者姜信圭坦言, AI的出现对于影视行业带来的改变如同工业革命, 但AI影视不可能完全取代传统创作, 因为演员身上独特的“气息”是AI无法复制的。

姜信圭认为, AI当然有其强大的表现力和优势, 但演员的呼吸、表演中传递的信息, 源自其走过的人生, 这是机器不可能拥有的。他相信, AI最终会成为电影制作中特效技术的一部分, 与传统电影并行发展。姜信圭表示, 未来他的电影创作将是“双轨”并行模式: 一方面, 继续使用传统方式拍摄在现实条件下能实现的电影; 另一方面, 利用AI去尝试那些原本想都不敢想的故事。

(新华社韩国釜山9月23日电)

换脸、代过、写代码 AI很“忙”, 别当法“盲”!

新华社记者 周闻韬 宋立崑

随着AI技术不断发展, 一些新型违法犯罪行为开始冒头, 给网络空间安全和群众人身财产安全带来威胁。记者采访多地公安机关, 揭示犯罪手法, 提升防范意识。

学AI, 竟为“换脸”行骗

“你们公众号怎么开始推荐投资理财App了? 靠谱吗?”今年6月10日, 某机构工作人员像往常一样打开公司公众号评论区, 却被一连串粉丝留言惊出了一身冷汗。工作人员发现, 其运营的公众号不知何时竟发文称即将停更, 并号召粉丝关注另一个投资理财类账号。工作人员尝试登录账号后台, 发现不仅密码被修改, 连公司法人代表信息都被篡改了。

意识到事态严重, 工作人员第一时间报案。这也是湖北省首起利用AI换脸技术非法侵入计算机系统案。

接警后, 武汉网警迅速成立专案组, 研判发现被盗公众号的操作痕迹为: 犯罪分子通过“AI换脸”

技术, 更换了公司法人信息, 又用新的“脸”识别登录该账号, 进而发布涉诈引流信息。

顺着线索追踪, 专案组很快锁定了远在山东潍坊的犯罪嫌疑人阿成(化名)。

站在民警眼前的阿成, 衣着朴素, 从事大棚种植, 一度令警察怀疑追错了人。但对其住所搜查时, 办案民警通过技术手段从其电脑已删除的电子数据中, 找到了大量AI换脸素材及非法所得的虚拟货币。

原来, 阿成本是美术技工, 嫌寻常做图收入不高, 就改行务农, 又动了做AI图挣“轻松钱”的歪脑筋。

2022年5月, 他在外网接触“人脸代过”灰色产业, 加入相关群组后, 先在各类小型电商平台接单制作人脸图像, 每张收费200元; 随着技术提升, 他掌握了生成动态人脸视频的方法, “报价”也水涨船高, 破解一张AI动态的人脸最高能卖到1000元。案发时, 已非法获利40余万元。

“跑马机”作弊成黑产

还有不法分子瞄上培训学时, 用AI帮人“打卡”。

今年3月, 重庆警方侦查发现, 一些驾校学员无需实际练车即可刷满学时, 背后是不法分子使用“跑马机”并结合AI技术实施作弊。这一犯罪链条涉及生产、销售、使用等多个环节, 已形成黑色产业。

什么是“跑马机”? 重庆市南岸区公安分局网安支队民警介绍, 这是利用汽车脉冲信号原理制作的设备, 其核心功能是通过侵入并篡改驾考培计时系统, 并运用AI技术模拟学员动态人像, 达成伪造培训记录目的。

通过涉案资金追溯显示, 利益链条的源头是驾校为了降低运营成本, 以提供“快速拿证”为噱头, 吸引学员大量报名缴费获利。与此同时, 上游负责销售“跑马机”的黑代理商也获利不菲。此外, 部分驾校还外接订单“代打卡”获利, 其规模化运作模式显示, 涉“跑马机”

犯罪已演变为机构化犯罪。

截至目前, 重庆警方已打掉涉案违法犯罪团伙2个, 抓获犯罪嫌疑人70名, 查扣“跑马机”设备384台, 查处涉案驾校34家。

AI写代码, 竟成了黑产源头

今年2月, 重庆万州区网安部门发现, 辖区犯罪嫌疑人王某针对某社交软件开发群控程序。经调查, 一个使用黑产软件从事各类违法犯罪的团伙浮出水面。

据万州区公安局网安支队民警介绍, 大专学历的王某自学掌握相关技术后, 开始招募多名同伙用AI技术编写程序代码, 制作各类黑产应用程序。这类黑产软件无需使用官方客户端, 即可直接与后台服务器进行数据交互, 具有“多开”“群控”“批量管理”等功能, 可实现批量发送消息、红包等操作。

令人惊讶的是, 王某还根据下游团伙的犯罪需求, “定制”出各种黑产软件, 成为赌博、网络水军、电信诈骗等多个犯罪链条的技术源头。

比如, 一到案的犯罪嫌疑人曾是电商从业者, 因觉得来钱慢, 便与王某技术对接, “转行”专门从事刷单水军活动, 其在城中村租下一套房子设立“工作室”, 招募多人加入。抓捕时, 警方在房间里当场查获多台用于作案的电脑和手机。

目前, 专案组已顺藤摸瓜, 先后打掉位于重庆、福建、江苏等地从事黑产软件开发及网络水军犯罪团伙4个, 抓获犯罪嫌疑人15名, 查获各类黑产程序软件25个, 查扣资金500余万元, 作案电脑、手机100余台。

AI时代已来临, 新技术能做的事越来越多, 也越来越有想象力。但筑牢安全底线的第一守则就是: AI可以很“忙”, 但使用它的人不能法盲。掌握AI技术的专业人员, 切勿因贪欲走上违法犯罪道路。同时各方应加强技术监管, 进一步完善生物特征检测等AI时代的防伪技术, 强化落实对个人隐私和信息的技术保护和法律责任。

(新华社重庆9月24日电)

讲文明 树新风 公益广告

坚持生态保护第一 守好蓝天碧水净土

福州西湖|陈暖 摄

中共福州市委文明办 宣