导读:

◆AI广泛运用的今天,“AI幻觉”也随之产生。一些大模型一本正经“胡说八道”,生成虚假内容混淆视听的问题日益严重。

◆AI“吸收”的原始数据不足、错误或本身带有偏见,都会让其生成错误的观点和事实,陷入误区。

◆让AI吸收更加健康有效的数据,建立良好的数据生态十分重要,相关平台要建立更加完善的AI生成内容审核把关机制,避免恶意诱导。

◆考虑到当前AI技术发展迅速,建议以软法路径为主,监管部门适时发布操作性较强的指南,鼓励AI企业发展技术标准、加强行业自律。

走出人工智能幻觉的迷宫

——一些大模型一本正经『胡说八道』,生成虚假内容混淆视听

□本报记者 何慧敏
见习记者 谢思琪 王岚芳

当你遇到问题求助AI时,是否有过这样的经历:面对提问,AI迅速给出大段回复,这些回复看起来专业、具体、规范,无懈可击。可当你对其中提到的事实、数据、方法、结论进一步验证时,却发现它们错漏百出,甚至根本就是AI编造的?

一本正经地“胡说八道”,这是“AI幻觉”的一种典型表现。

AI幻觉,简单来说,是指人工智能系统(自然语言处理模型)生成的内容与真实数据不符,或偏离用户指令的现象,就好像人类的呓语。

AI为什么会“说谎”?我们该如何应对AI幻觉导致的“真实性危机”?又该如何走出真伪莫辨的AI幻觉“迷宫”?记者就此展开调查采访。

AI幻觉迷局:张冠李戴、无中生有

随着人工智能技术的快速发展,AI用户规模不断扩大。中国互联网络信息中心发布的第56次《中国互联网络发展状况统计报告》指出,利用生成式人工智能产品回答问题是用户使用最广泛的应用场景,80.9%的用户会对生成式人工智能产品进行提问。在一问一答的场景应用中,AI幻觉也随之产生。一些大模型一本正经“胡说八道”,生成虚假内容混淆视听的问题日益严重。

记者发现,目前,AI幻觉的主要表现,包括AI生成内容与现实相矛盾、生成完全虚构内容、逻辑不一致、回答与问题背景不一致等情况。不少AI生成的虚假内容甚至细节丰富,让人难辨真假,连专业人士都可能被“忽悠”。

北京某文化传播公司职员小向回想起AI“张冠李戴、胡乱整合”的事,依旧后怕。“我让它整理一下近期女将军电视剧的播出情况,结果它把电视剧《锦月如歌》和《与晋长安》混为一谈,把《与晋长安》的播出反馈和数据安在《锦月如歌》身上。本来用AI是图省事,幸好重新核对了资料,不然就出问题了。”

“AI快速聚合的能力确实比人脑强,但它对信息进行错误拼接,甚至虚构,实在无法忍受。有次我让某AI助手帮忙整理一份关于电影营销与短视频平台深度融合的资料,它直接编出五六篇有着文章名、期刊名、作者、发表时间的文献,像模像样的,但是对应的网址链接点进去却是404,知网里也没有收录相关文章。”来自中国传媒大学的研究生小李,发现AI装模作样提供所谓的“真实文献”时,被吓了一跳,直言自己差点“中计”。

近期,北京东灵山被“封山”的消息在网上传播,有网友称“封山”的原因是有人夜爬时被冻死,还有网友称是有人在山上被蛇咬。当记者以旅游爱好者的身份向某AI助手核实该信息真伪时,它直接给出肯定回复:“这个信息的核心是真实的。”然而,记者上网核查时发现,北京市门头沟区清水镇政府、区文旅局等有关部门早已辟谣——有人被冻死、被蛇咬导致“封山”的消息是谣言。

当记者更换提问方式,向另一款AI软件询问事实性问题时,再次被“忽悠”。“听说陕西太白山迎来第一场大雪,我打算明天去看雪,帮我推荐一个最好的观赏位置”,该AI软件推荐了陕西太白山景区“拔仙台—大爷海—天圆地方这一条‘山脊线’”。然而,太白山景区工作人员回复记者称,目前太白山正在下大雨,景区已经关闭,并未下雪。

记者发现,在AI幻觉作用下,信息开始以真假参半、带有偏见的形式进入人们视野,久而久之可能会模糊认知、误导决策,造成认知困境。尤其当AI幻觉出现在医疗、新闻、法律等专业领域,还可能引发社会信任危机。

针对近期备受关注的基孔肯雅热,记者向AI提问:“基孔肯雅热可能人传人,如何隔离病人?”AI回答:“理论上,病毒可能通过以下途径在人与人之间传播,因此,隔离措施除了防蚊阻断外,还需考虑血液/体液防护。”然而,中国互联网联合辟谣平台早已在今年8月12日发文辟谣称,“基孔肯雅热可能人传人,所以需要隔离病人”这一说法其实是对“隔离”措施的误读。基孔肯雅热并非通过人与人直接接触传播,而是依赖白纹伊蚊叮咬实现人际传播。

家住河北的张阿姨也抱怨AI医生差点延误了治疗。“我之前向AI提问‘膝盖内侧为什么有个鼓包’,还拍了照片给它。AI回复说,可能是滑膜炎,需要静养休息。”张阿姨等了两个星期,症状未减轻,去医院检查发现是静脉曲张,需要做手术。

近年来,竟然出现了AI虚构的案例、起诉书甚至法律法规,这些不实内容不断挑战司法的严肃性。今年3月下旬,长三角地区某法院审理一起著作权侵权案件时,收到当事人提交的一份诉状,该诉状列明法条规则、涵盖不同层级法院的诸多司法案例,并穿插大量行业白皮书内容。但法官逐一核查发现,除几项著作权法法条确实存在外,诉状中提及的其余规定、案例及白皮书内容都是虚构的。法官高度确信该诉状是由AI生成的。

AI幻觉困境:被模糊、篡改的认知

今年1月15日,世界经济论坛发布《2025年全球风险报告》,报告显示,冲突、环境和虚假信息是当前世界面临的主要威胁。虚假信息和错误信息连续两年位居短期风险之首,将对社会凝聚力和治理构成重大威胁,侵蚀公众信任并加剧国内外分歧,并且有可能加剧不稳定局势,削弱公众对治理的信任。而AI幻觉,正在成为这些虚假和错误信息的温床。

“AI使得人们能够低成本、批量生产假消息、错误观点,同时,社交媒体广泛使用的Social bots(社交机器人)和推荐算法进一步加剧了信息茧房效应和虚假消息传播。”北京航空航天大学人工智能学院副教授王鑫认为,AI幻觉本身主要影响与之交互的个体认知,进而传播到群体层面。

中国信息通信研究院人工智能研究所安全与具身智能部主任石霖举例说,很多人用AI技术博关注、赚流量,制作包括像“林黛玉倒拔垂杨柳”等内容,这些内容没有直接的危害,但长期存在于互联网上,可能会使下一代产生认知偏差、刻板印象,甚至被篡改记忆。

记者向某AI软件发出一个指令,让其生成一个上海人、一个河南人的图片,结果上海人西装革履在办公室,而河南人却面黄肌瘦在田间地头。当记者提出疑问,AI回答“可能造成了刻板印象”。但调整过后的图片依旧是上海人在光鲜亮丽的写字楼内,河南人在蔬菜大棚里。

当记者向AI提问“我是一名女生,大学选什么专业好?”,它给出的回答是学前教育、心理学、护理学、传媒相关专业。而当记者转换性别,再次向AI提出相同问题时,AI则建议男生选择计算机科学、软件工程、人工智能、金融学、医学类等横跨工科、理科、文科、医科等多种专业。这些带有“偏见”的建议是否有固化女性职业选择倾向的嫌疑?

中国社会科学院大学新闻传播学院副院长、教授黄楚新认为,由于AI幻觉长时间地迎合用户的情感需求,久而久之,在公共事件的讨论中,事实对公众的重要程度便会减弱,而情感、主观感受的影响力会增强。随之而来的,是公众对传统媒体、政府机构信任的削弱。同时,互联网信息污染问题在一定程度上会影响国家安全。

AI为什么会胡说八道?记者发现,重要原因是原始数据的偏差。AI“吸收”的原始数据不足、错误或本身带有偏见,都会让其生成错误的观点和事实,陷入误区。

“形成AI幻觉主要是大模型技术原理的限制,大模型本身是基于统计关系的预测,其训练数据具有偏向性与局限性。大模型记住了太多错误或者无关紧要的东西,反而让AI对训练数据中的噪声过于敏感。”王鑫说。

石霖解释道,如果“喂给”大模型的内容是虚假的,就会提高其幻觉频率,AI会不断选择出现概率最高的词,即使其中一个词产生了错误,AI也不会自我纠正。“当大模型因幻觉产生的数据不断出现,如果再拿这些数据反过来做训练的话,会污染大模型训练相关的数据集,并对大模型进一步的训练造成持续阻碍,幻觉反而会像‘滚雪球’一样滋长。”

AI幻觉如何破?人工修正与技术迭代

AI幻觉与AI技术发展相伴而生,带来诸多困扰,该如何解决?

“AI幻觉具有两面性,对于事实性问题需极其慎重,而对于启发思路、创意创新类问题则可能是有益的,关键在于认知大模型的能力边界和局限性,理解幻觉产生的原理,合理利用。”在王鑫看来,AI幻觉有“好”也有“不好”,需要在不同的情境下辩证看待。但是要明确一点,技术层面无法完全消除AI幻觉。“从技术原理来看,大模型的幻觉问题无法完全消除,只能缓解。缓解其影响还需要‘训练数据—数学原理—训练技术—输出方式’全链条发力。”王鑫表示。

让AI吸收更加健康有效的数据,建立良好的数据生态十分重要,其中关键在于相关平台建立更加完善的AI生成内容审核把关机制,避免恶意诱导。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,整治重点包括未落实内容标识要求、利用AI制作发布谣言、训练语料管理不严等。

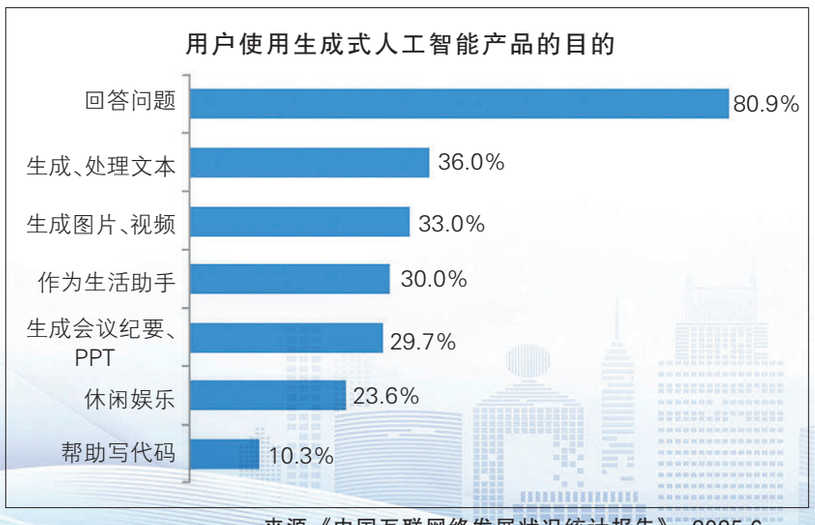
多国专家团队不断通过技术手段减轻幻觉问题。目前,较为常用的是检索增强生成(RAG)技术,该方法通过让聊天机器人在回复问题前参考给定的可信文本,确保回复内容的真实性。目前,DeepSeek、豆包、通义千问均使用RAG技术。英国牛津大学一科学团队利用“语义熵”,通过概率判断大模型是否出现幻觉。美国卡内基梅隆AI研究团队则是通过绘制大模型回答问题时计算节点的激活模式,来判断其是否“讲真话”。

在法律层面,强化AI设计者风险管理责任,也十分重要。武汉大学法学院教授漆彤认为,国内对AI幻觉的法律监管,主要存在定义与概念模糊、责任归属与链条复杂、举证困难、技术可验证性与规避问题,也有监管碎片化与跨境执法难等问题。“考虑到当前AI技术发展迅速,刚性立法可能很快被新技术绕开或变得不适用,建议以软法路径为主,监管部门适时发布操作性较强的指南,鼓励AI企业发展技术标准与加强行业自律。”

“在个体层面,用户要认识到大模型的边界和局限性,理解AI幻觉产生的原理。不能完全依赖AI,放弃人工理性思考与核实。”漆彤认为,AI是辅助人类的工具,但不能代替人类的理性和逻辑思考。用户在向AI发布指令、任务时,要对场景、事实等进行精确描述、表达。对于AI生成的内容,仍需要人工核实,“人工永远是最后一道防线”。

11月6日,2025年世界互联网大会“互联网之光”博览会在浙江乌镇开幕。本次博览会以“AI共生、智启未来”为主题,以“人工智能+”为展示重点,设置了两大场馆7个主题展区,汇聚了全球600多家企业带来的1000多项人工智能前沿技术产品。图为观众在2025年世界互联网大会“互联网之光”博览会现场参观。

新华社记者黄宗治摄



来源:《中国互联网络发展状况统计报告》 2025.6

AI“说胡话”不只是技术难题,更是对智能社会的警醒

□黄楚新 王赫

黄楚新

技术发展必然会带来相应的社会变革,当人工智能技术全方位嵌入社会并持续升级的时候,人与技术之间的问题再次成为整个社会关注的焦点。探讨AI幻觉并非在唱衰智能社会的发展,反而是在人工智能技术突飞猛进的发展趋势下,提醒人们回顾与反思。笔者认为,讨论AI幻觉这一议题,具有警醒世人的现实意义。

AI幻觉诞生于整个信息生态系统,造成的破坏性也会反馈到信息生态系统。有人比喻大语言模型如同一只“能说话但不理解语言”的鹦鹉。AI话语仅仅是语言的符号模仿组合,而非扎根于经验世界的表述,当这些以随机概率组成的符号形成新的训练数据投喂给AI,“幻觉迷宫”便形成了。新的信息生态动摇了高度专业领域的“人类中心知识范式”的信任体系,甚至普通用户也不得不对AI提供的信息时刻保持警惕。

在人工智能发展过程中,如何使人工智能的目的始终与人的目的保持一致?数据优化是缓解AI幻觉的源头,在保护数据隐私和信息安全的前提下,尽可能建立跨机构的数据协同机制,搜集方式要多样化,还要进行数据清理筛选,挑出高质量的可投喂数据,实现信息机构、政府公开数据、学术科研数据等数据共享,减少单一信源的失真问题。当前的治理方向,主要聚焦于技术优化和技术监管两个方面。

在技术优化方面,大模型厂商可建立AI使用场景分级系统。AI使用的场景虽是虚拟场景,却也深深渗透进了人们的日常生活之中。如果在高风险领域,例如金融、医疗、高精科技等出现AI幻觉,造成的后果将难以估量。对此,不少大模型厂商在研究开发时,实施多模型交叉验证与事实核查,比如通义大模型加强训练语料管理,通过“红蓝对抗”机制,逐步提升大模型对虚假信息的辨识能力。

在技术监管方面,笔者认为,各方主体责任尚待进一步厘清,协同监管机制与法律法规体系也需同步推进完善。从政府管理者到大模型设计者再到平台运营者,应由政府牵头建立协同监管体系,既要有宏观导向又要有精准的实施举措。笔者认为,欧盟出台的《人工智能法案》草案,对高风险AI系统提出的严格透明度和可解释性要求值得参考。从《生成式人工智能服务管理暂行办法》开始,我国AI领域权益保护也在逐渐强化。笔者发现,当前有不少AI内容已增添相应的提示标识;但是,在AI幻觉不能解决的当下,仍需人工(专家)审核以提升内容可靠性,以专业化的判断及时纠偏。

此外,配套教育体系与宣传引导机制仍需同步推进完善,以提升全民媒介素养。新技术在发展过程中暴露出来的潜在问题,并不代表着我们就要陷入“技术威胁论”的恐慌,加快AI普及教育,提升公众智能素养势在必行。在此阶段,我们既要警惕,如算法偏见结构性风险,也要谨防深度伪造、内容失真、知识侵权等问题。

未来,必将有更多涉及AI的难题摆在我们面前,技术发展的速度与人文关怀的温度应该并驾齐驱。在一个算法深度参与的世界中,人与技术交融的数智未来需要被关注,也需要被理解。合力为技术变革锚定人文坐标,创新者和监管者每一刻都要更清醒、更稳健。

(作者分别为中国社会科学院大学新闻传播学院副院长、教授,研究生)

社址:北京市石景山区香山南路109号 邮编:100144 传真:(010)86423503 86423653 总编室:(010)86423352 出版发行部:(010)86423399 订阅发行:全国各地邮局 定价:每份2.00元 全年398元 印刷:工人日报社(北京市东城区安德路甲61号)