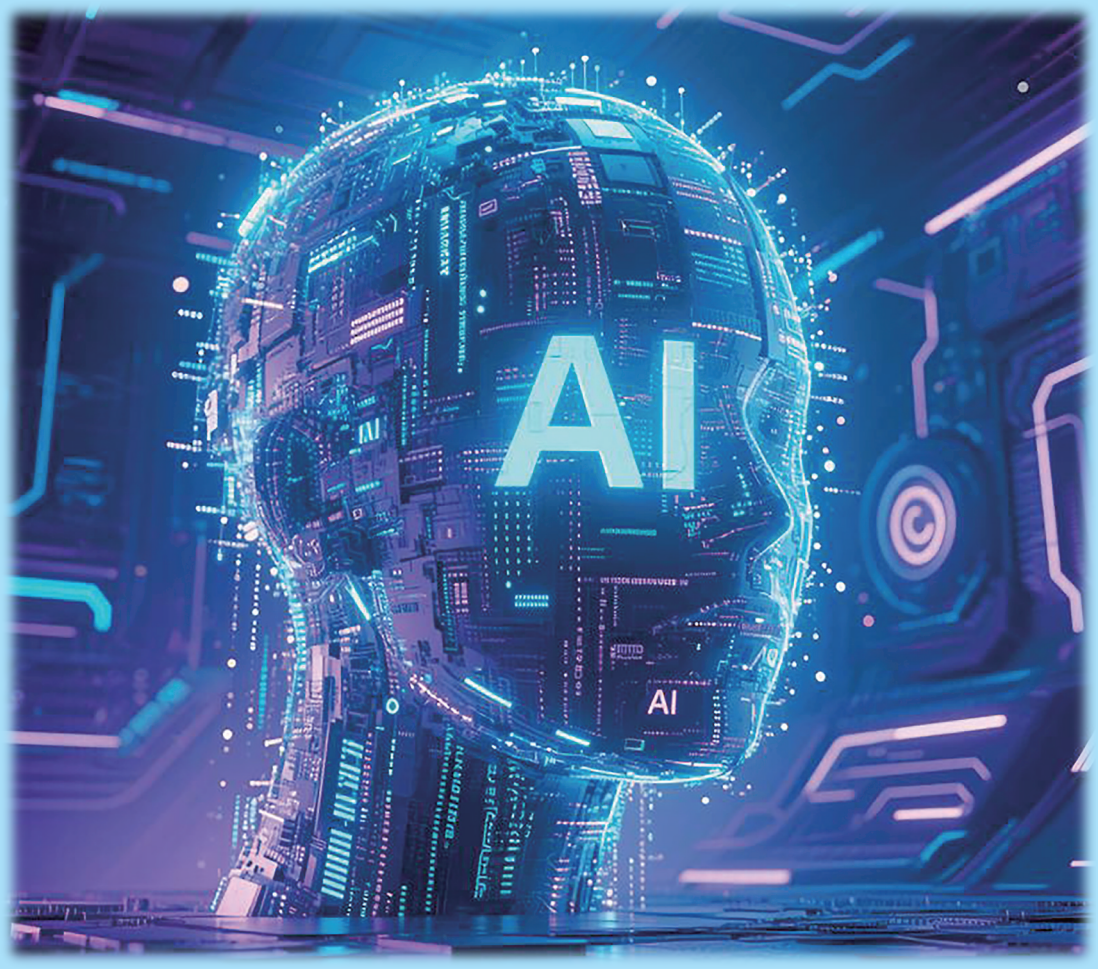
AI幻觉频现 风险挑战几何

当前,人工智能技术快速发展,但大模型“自说自话”、一本正经“胡说八道”、生成偏离事实内容的问题日益凸显,这一现象被称为“AI幻觉”。不少业内人士提醒,由于大模型主要基于概率生成文本而非逻辑推理,在短期内难以完全避免此类问题。

《瞭望》新闻周刊记者观察到,AI虚构事实或逻辑混乱的“幻觉”已在法律、内容创作、专业咨询等多个领域造成实际影响。

AI技术的发展方兴未艾,但确保其生成内容的真实性和可靠性,尤其需要技术开发者、监管机构共同努力。针对“AI幻觉”问题,业界建议,在技术层面,要持续优化模型架构,增强事实核查能力;在监管层面,需完善相关规范,明确责任边界。



幻觉频现

“AI幻觉”已经成为当前AI技术发展中最突出的技术瓶颈之一。从技术层面来看,AI幻觉的产生主要源于三个方面的原因:首先是训练数据的不足或偏差,导致模型对某些领域的认知存在缺陷;其次是算法架构的局限性,当前主流大模型主要基于概率预测而非逻辑推理;最后是训练目标的设定问题,模型往往更倾向于生成“流畅”而非“准确”的内容。

“AI幻觉主要表现为事实性幻觉和逻辑性幻觉两种。”业内人士介绍,事实性幻觉,表现为模型会编造完全不存在的事实或信息。这种情况在专业领域尤为突出,如在法律咨询中虚构判例,在医疗诊断中给出错误结论,或杜撰历史事件。逻辑性幻觉,表现为模型在长文本生成或连续对话中出现前后矛盾、逻辑混乱的情况,这主要是因为当前大模型的注意力机制在处理复杂语境时存在局限性。

多项研究证实了AI幻觉问题的严重性。今年3月,哥伦比亚大学数字新闻研究中心针对主流AI搜索工具进行的专项测试发现,这些工具在新闻引用方面的平均错误率达到60%。一些研究显示,AI并不擅长辨别“新闻事实来自哪里”,会出现混淆信息来源、提供失效链接等问题。更令人担忧的是,随着模型规模的扩大,某些类型的幻觉问题不仅没有改善,反而呈现加剧趋势。

多位业内专家认为,AI幻觉问题在现有技术框架下难以彻底解决。思谋科技联合创始人刘枢表示,当前的大模型架构决定了其本质上是一个“黑箱”系统,优化结构等方式只能缓解模型幻觉问题,很难完全避免幻觉的产生。

有受访者从认知科学的角度分析,当前的大模型处于“我不知道我知道什么”的状态,缺乏对自身知识边界的准确判断能力。这些技术特性决定了AI幻觉问题的存在,需通过多方面的技术改进来逐步缓解。

警惕风险

业界普遍认为,在AI幻觉短期内难以完全消除的背景下,其潜在风险已从信息领域蔓延至现实世界,可能带来较大风险。

世界经济论坛《2025年全球风险报告》已将“错误和虚假信息”列为全球五大风险之一,其中AI生成的幻觉内容被视作关键诱因之一。

AI幻觉最直接的危害是造成“信息污染”。在法律领域,美国纽约南区联邦法院在审理一起航空事故诉讼时发现,原告律师提交的法律文书中引用了ChatGPT生成的6个虚假判例,这些虚构案例包括完整的案件名称、案卷号及法官意见,甚至模仿了美国联邦最高法院的判例风格,其虚构能力干扰了司法程序。

金融咨询领域,AI可能给出错误投资建议,如误读财报数据或虚构企业信息。

更令人担忧的是,这些错误信息可能被其他AI系统吸收,形成“幻觉循环”——错误数据不断被强化,最终污染整个信息生态。

随着AI技术向实体设备领域渗透,幻觉问题的影响已超越虚拟范畴,可能对人身安全构成威胁。在自动驾驶领域,生成式AI被用于实时路况分析和决策制定。业内人士表示,在复杂路况中,自动驾驶若产生“感知幻觉”,可能导致系统误判环境,触发错误决策,直接威胁行车安全。

人形机器人领域风险更值得关注。优必选副总裁庞建新说:“当机器人因幻觉做出错误动作时,后果远超文本错误。”例如,护理机器人可能误解指令给患者错误用药,工业机器人可能误判操作参数造成生产事故。这些场景中,AI幻觉甚至可能威胁人身安全。

协同治理

对于AI幻觉问题带来的挑战,业内人士建议从技术创新、制度监管等多个维度构建综合治理体系。

技术创新是解决AI幻觉问题的根本途径。近年来,业界已提出多种技术方案来应对这一挑战。刘枢等介绍,检索增强生成(Retrieval-augmented Generation,RAG)技术融合了检索与生成模型优势,是当前重要的发展方向之一。其通过将大模型与权威知识库实时对接,能显著提升生成内容准确性。

全国政协委员、360集团创始人周鸿祎提出“以模制模”,构建专业知识库,实施合理的矫正机制,构建更完善的安全防护体系,降低“幻觉”带来的负面影响。例如,研发专用的安全大模型来监督知识库使用和智能体调用,并采用多模型交叉验证、搜索矫正等技术手段来识别和纠正幻觉。

制度监管方面需要建立多层次的治理体系。云天励飞品牌运营中心总经理胡思幸认为,要完善监管治理,研究AI生成内容“数字水印+风险提示”双重标识,为AI生成内容提供有效的溯源和警示机制。针对日益突出的AI造谣问题,法律界人士建议,要持续完善相关规定,明确利用AI造谣的法律责任,加大对违法行为的惩处力度。

治理体系之外,当前阶段,在用户使用AI开展工作时,亦需要建立对AI能力的理性认知,了解其局限性。培养多渠道验证信息的习惯,优先选择权威、可信赖的媒体或机构作为信息来源,这些基本素养的提升将有效降低AI幻觉的社会影响。同时,在与AI系统交互时应保持必要的怀疑态度和批判思维,多渠道核查验证信息的准确性。

《瞭望》新闻周刊记者 孙飞 陈宇轩

印刷:茂名日报印务有限公司 地址:茂名市电白区(原茂港区)进港大道128号 广告热线:2963992 2963993 茂名日报社广告发布登记通知书编号:440900100009