## 寻医问药、AI合成、人脸识别、内容创作——

# 人工智能安全问题存隐忧

4日,全国人大代表雷军公布了其准备的“2025两会建议”。其中一条,是关于加强“AI换脸拟声”违法侵权重灾区治理。

随着人工智能被广泛应用,随之而来的安全问题也愈发凸显。记者收集了四大安全问题——寻医问药、人工智能合成、人脸识别、内容创作,请中国网络空间安全协会人工智能安全治理专业委员会(以下简称“专委会”)三位专家委员韩蒙、周杨、王晓梅解答。

### 拿着“AI诊断”寻医问药?精准判断还得靠人

“患者拿着DeepSeek来看病,还质疑我过度检查……”越来越多的患者拿着AI给出的诊断,与医生争论。

“患者用DeepSeek寻医问药,不能说完全不对,但不能只依赖于它的判断。”杭州医策科技有限公司创始人王晓梅告诉记者,目前像DeepSeek、ChatGPT都属于人工智能的大语言模型产品,帮助人机实现语言和文字的交互。这类产品可以触达所有公共数据。在医学方面,它们可以触达公开的医典、论文等。然后,基于这些给患者一个回答。“这个回答是通识性的。”各个专业领域还有高质量私域数据,医学上比如影像、病理、基因等,它触达不到。

而且,医学上有很多非语言、非文本的数据,比如每个人都很熟悉的医学影像。“需要多模态大模型,用文本、图像、视频、音频等多模态信息联合训练。”王晓梅说,患者去医院看病,医生会依据病情进行相关检查,得到血液、生化、影像、病理、基因等检测报告(数据),再依据报告给出诊疗方案。“要达到对疾病的精准判断,确保安全性和有效性,就需要有医疗器械注册证加持的诊疗级人工智能产品。”

近来,也有越来越多的医院在应用人工智能技术。医策科技的病理人工智能辅助诊断系统就是诊疗级人工智能产品,在获批当地药监局颁发的医疗器械注册证后,已在国内外多家医疗机构部署应用,辅助医生判读达几百万例病理报告。“中国每百万人的病理医生数量仅为1.5人,AI在病理分析上的应用,可以大大减轻病理医生不足这个压力。”

### 用AI生成别人的脸和声音?可以置入隐式标识

最近,雷军、刘晓庆、刘德华、张文宏等名人,纷纷被人工智能技术换脸拟声合成视频。这类视频以假乱真,很容易“欺骗”受众。

北京市竞天公诚律师事务所合伙人周杨告诉记者,这类用AI合成的名人音视频,会侵犯名人的肖像和声誉,还存在散布虚假、不良信息的风险。她举例:“DeepSeek爆火之后,《黑神话:悟空》制作人冯骥发文盛赞。很快,一条DeepSeek创始人梁文峰的回应,引发朋友圈广泛转载。最后,多个知名大V声称梁文峰的回应为AI伪造。”

AIGC(人工智能生成内容)与过去影响网络空间安全的内容一样,都会出现虚假和不良信息,但更加难以甄别,更容易导致不良影响和后果。《互联网信息服务管理办法》明确规定了网络发布信息“九不准”。《网络信息内容生态治理规定》还有“十一不得”。国家互联网信息办公室还发布了《人工智能生成合成内容标识办法(征求意见稿)》,与全国网络安全标准委员会发布的《网络安全技术 人工智能生成合成内容标识方法》强制性国家标准(征求意见稿)配套使用,要求对人工智能生成合成内容进行标识,规范AIGC在网络空间中的流通安全。

此外,还有合成普通人音视频、对其亲友实施诈骗的行为,未成年人和中老年尤其容易被针对。周杨表示,《网络信息内容生态治理规定》《生成式人工智能服务管理暂行办法》《未成年人网络保护办法》均有关于弱势群体的专门规定。

浙江大学研究员韩蒙表示,AI合成伪造和AI标识,存在相互博弈的关系。在技术上,可以对AIGC进行标识。“标识有显式的,比如现在大家在视频网站上看到的水印。随着技术升级,包括我们团队在内,一些国内团队也对隐式标识展开研究。”当利用AI生成作品时,隐式标识会被藏在作品中,肉眼是看不见的。但通过检测手段,标识可以显现。

AI复刻的模特脸。 CFP供图

### 戴着头套骗过人脸识别?依赖场景分级应对

网络上有声称“可以攻破人脸识别系统”的硅胶头套售卖。

周杨介绍,利用硅胶头套导致人脸识别系统识别通过,本质上是一种对人脸识别技术及系统的攻击行为,考验的是人脸识别技术和系统的安全性问题。关于人脸识别技术的应用安全,国家网信办曾发布《人脸识别技术应用安全管理规定(试行)(征求意见稿)》,从应用层安全、技术层安全、数据层安全等多个层面规定了人脸识别技术的提供方和应用方应当承担的相关责任。

“我们都希望人脸识别系统既轻松又精确,我们戴个帽子、口罩,换个发型,它都能把我们识别出来。”但是,韩蒙介绍,AI模型在人脸识别泛化性和精准性方面,一直像跷跷板,“不可能提供完美泛化性的同时,又提供完美精准性。”

对此,在使用场景上进行分级分类是个相对不错的应对方案。比如:上班打卡等场景,可以对精确性要求低一点;金融转账等高度安全需求场景,精确性要求必须高。“所以,现在有活体识别技术,让你张张嘴、眨眨眼等。”

央视曾用硅胶头套做过测试,结果显示头部厂商的增强版人脸核实身份技术,能准确识别和有效拦截攻击。

### 用AI帮着自己搞“创作”?独创才是作品灵魂

近日,武汉东湖新技术开发区人民法院科学城法庭审结一起“AI生成图被侵权”的著作权纠纷案。AIGC创作者王某,将盗用其图片作品的某公司诉至法院。法院认为该公司应当承担停止侵权和赔偿损失的侵权责任,赔偿王某经济损失及合理开支4000元。

“AIGC的出现模糊了创作的主体边界。”承办法官表示,AIGC是否构成作品需要个案判断,核心在于作者的智力投入是否达到独创性标准以及生成作品是否高度呈现了作者的独创性表达。

周杨也认为,如果创作者只是给AI提供几个简单的关键词,那很难认定其有作品的著作权。“有独立的逻辑,不断地创作、调整、筛选,有大量的个人特色和智力付出在作品里,才能具备产生著作权这一保护人类智力创作权利的条件。”

大学生利用AI来完成论文撰写,周杨认为,把AI当成一种先进生产力工具无可厚非。但仍应遵循法律法规的规定,不得输出包含大量抄袭、改编等侵犯他人知识产权的作品,而应当加入自身的观点、思考、创造力,形成真正有保护价值的论文。此外,AI模型在训练时,也应当考虑此类伦理问题,加强训练数据的管理和输出物的标识,在输出物合法合规且合乎伦理道德方面进行技术控制。

对于AI的管控,目前我国已经从输入端(训练数据),模型端(模型安全、伦理安全等)以及输出端(生成物内容安全)等方面进行了部署,并制定了相关落地细则和标准。“AI本质上是工具,怎么用、用来干什么,最终仍将取决于一个人自身的道德、目标与决策。”

据北京晚报