标识,让AI生产的内容“亮明身份”

■ 光明日报记者 **刘坤** 通讯员 **柳素雯**

夜深人静时,有人找它线上聊天、获取慰藉;视频带货时,有人用它快速生成营销文案;同学聚会时,有人用它现场作诗助兴……它,就是很多人眼中的“生活好帮手”“工作好搭子”——人工智能(AI)大模型。

近期,多款大模型产品横空出世、风靡全球,再次掀起人工智能热潮。但与此同时,也有一些人利用人工智能生成合成虚假内容,或为吸引眼球、赚取流量,或试图以假乱真、造谣诈骗等,带来一定风险和隐患,引发社会关注。

近日,国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局联合发布《人工智能生成合成内容标识办法》(以下简称《办法》)。《办法》聚焦人工智能“生成合成内容标识”关键点,通过标识提醒用户辨别虚假信息,明确相关服务主体的标识责任义务,规范内容制作、传播各环节标识行为,将于2025年9月1日起施行。

如何让人工智能生成合成内容“亮明身份”,不再“真假难辨”?如何破解人工智能安全治理难题?记者就此进行了采访。

人工智能合成让人真假难辨

小猫小狗随着音乐跳起舞蹈,家中“萌娃”大战公鸡……在短视频平台上,人们不时会刷到类似视频。其实,一些传播量、点赞量都很大的“神奇视频”并非是真实的,而是由人工智能生成合成的。

所谓人工智能生成合成内容,是指利用人工智能技术生成、合成的文本、图片、音频、视频、虚拟场景等信息。

近年来,人工智能技术快速发展,为生成合成文字、图片、音频、视频等信息提供了便利工具,海量信息得以快速生成合成并在网络平台传播,在促进经济发展、丰富网上内容、便利公众生活的同时,也带来生成合成技术滥用、破坏网络生态等问题,引发社会各界的关注。

“一些喜欢我参演的影视剧的观众,被AI换脸视频骗得很惨,这个性质非常恶劣。”一位演员说,希望能建立更好的规范。实际上,不少公众人物都遭遇过类似问题。

公安部数据显示,近两年来,全国共发生“AI换脸”类诈骗案近百起,累计造成的经济损失高达2亿元。

“人工智能生成合成内容日益逼真,也催生虚假消息传播、身份信息冒充、恶意内容生成等新型安全风险,并削弱着公众对网络传播内容的信任根基。”中国工程院院士、浙江大学教授陈纯说。

中国电子信息产业发展研究院副总工程师刘权认为,人工智能大模型本身存在不可解释性问题。大模型运作机制不透明,产生的“黑箱”属性导致不可解释性。这种不可解释性降低了可信度,使得输出的内容可能存在事实性错误和偏差,特别是在医疗、金融等领域,难以满足严苛的可信要求。

中国社会科学院大学互联网法治研究中心主任刘晓春表示,通过人工智能生成合成技术降低了内容“伪造”“造假”等的时间成本,尤其在图片、音频、视频等领域,从早期引起广泛关注的“换脸”工具,到伪造人声、通话视频的诈骗活动,再到“被压在废墟下的小男孩”等热点事件,人工智能技术带来的深度伪造、虚假信息、不良信息、抄袭侵权等问题,对包括内容治理在内的网络空间秩序构成挑战,对治理工具、方式和手段也提出了更新更高的要求。

数据显示,截至2024年底,共302款生成式人工智能服务在国家网信办完成备案,其中当年新增238款备案。

“当前,大模型能够生成高真实感的文本、人像、场景、音频,普通民众不借助检测工具往往很难辨别内容真伪。”中国科学院计算技术研究所数字内容合成与伪造检测实验室主任、研究员曹娟认为,随着生成式人工智能技术在各个行业落地应用,生成合成内容的数据规模快速增长,对生成内容进行全面检测成本太高。生成合成内容适用场景大幅扩展,助长了欺骗场景多样化。

打造可信赖的AI技术

《办法》的推出是我国推进人工智能领域安全治理、促进产业规范健康发展、引导技术向善的重要举措,标志着我国在生成式人工智能领域迈出了构建安全可信生态的关键一步。

国家互联网信息办公室有关负责人介绍,《办法》重点解决“哪些是生成的”“谁生成的”“从哪里生成的”等问题,推动由生成到传播各环节的全流程安全管理,力争打造可信赖的人工智能技术。

针对服务提供者,《办法》明确,应当对文本、音频、图片、视频、虚拟场景等生成合成内容添加显式标识,在提供生成合成内容下载、复制、导出等功能时,应当确保文件中含有满足要求的显式标识;应当在生成合成内容的文件元数据中添加隐式标识。

针对互联网应用程序分发平台,《办法》提出在应用程序上架或者上线审核时,应当要求互联网应用程序服务提供者说明是否提供人工智能生成合成服务,并核验其生成合成内容标识相关材料。

针对用户,《办法》提出,用户使用网络信息内容传播服务发布生成合成内容的,应当主动声明并使用服务提供者提供的标识功能进行标识。

此外,任何组织和个人不得恶意删除、篡改、伪造、隐匿本办法规定的生成合成内容标识,不得为他人实施上述恶意行为提供工具或者服务,不得通过不正当标识手段损害他人合法权益。

“《办法》以合理成本提高安全性,促进人工智能在文本对话、内容制作、辅助设计等各应用场景加快落地,同时减轻人工智能生成合成技术滥用危害,防范利用人工智能技术制作传播虚假信息等风险行为,推动人工智能健康有序发展。”该负责人说。

中国政法大学数据法治研究院教授张凌寒认为,人工智能生成内容标识制度具备独特制度价值。标识能够有效区分出人工智能生成合成的信息,防范虚假信息的传播利用;标识能够帮助用户快速了解生成式人工智能产品和服务的相关属性或参数信息;标识能够协助监管部门对生成式人工智能产品或服务实施评估、追溯等监管,推进人工智能生成合成内容合法合规发展。

张凌寒表示,目前,人工智能生成合成内容标识制度主要关注“是否机器生成”的形式判断。随着标识技术的进步与发展,未来标识制度能够发挥更为丰富的功能,从形式判断逐渐转向“是否足够可靠”的质量判断,进一步促进行业健康发展。

安全治理进入“深水区”

良法善治,重在实施。

为推动《办法》落地实施,强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》同步发布,更好地指导相关主体规范开展标识活动。

“当前,人工智能安全治理已从危害探讨进入实际执法的深水区。”曹娟说,大模型生成技术迭代更新很快,平均两个月就会出现新的里程碑模型,提升针对新生伪造方法的泛化能力至关重要,需要摒弃“来一个打一枪”的事后思维,构建AI鉴伪底座模型,提高鉴伪模型的泛化适用性能。

曹娟表示,要研究精准化对抗鉴伪技术,防范恶意逃避风险。同时,降低对无害生成内容传播的影响,兼顾生成内容应用的发展与治理,打造全民化伪造检测工具,推动人人可用鉴伪。

“可以预见,我国人工智能安全执法将延续重点事项监管、促进产业有序发展的方向。”公安部第三研究所副所长金波说,伴随标识管理与算法备案、安全评估等机制逐步实现有机衔接,生成合成内容标识合规或将成为相关部门开展人工智能监督检查、专项行动的重点关注领域。在此进程中,如何平衡好发展与安全、创新与责任,提升执法的专业化、精细化、智慧化水平,从而培育出安全、开放、公平的人工智能产业生态环境,还需要深入探究。

“此外,‘人的因素’应融入人工智能标识管理的全过程。”金波表示,要着重提升公众对信息内容真实性、来源可追溯性的评估能力,积极培育公众的人工智能素养,确保人工智能技术成果普惠共享。

在陈纯看来,《办法》的落地离不开全社会凝心聚力和协同配合。地方主管部门、高校、科研机构、企业可共同参与,形成标识工作先行的多点协同内容治理网络,推动标识工作行稳致远。此外,有必要建立全国性的内容标识公共服务平台,以可视可交互的实际操练形式,促进公众和产业深入理解标识工作。

“生成式人工智能促使信息技术从普通工具属性向‘思维伙伴’型高智能化工具属性进化,带来颠覆性发展机遇,未来势必成为我们的必备技能,这要求我们在使用时切实正确理解并管理人工智能。”陈纯说。

转自《光明日报》(2025年03月27日15版)

AI时代,如何让治理跟上技术步伐

■ 新华社“新华视点”记者 宋佳 王存福 刘开雄

从“AI儿科医生”参与病情会诊,到“AI数智员工”辅助提供公共服务,再到人形机器人进厂上班……全球人工智能技术进入爆发式发展阶段,正以前所未有的速度走入千家万户、改变千行百业。当AI重塑我们熟悉的世界,随之而来的机遇与挑战如何应对?应用与治理如何平衡推进?

博鳌亚洲论坛2025年年会上,国内外技术前沿领域专家、学者和企业代表就相关话题展开热烈讨论。

适应AI“新常态”

当下,人工智能加速赋能千行百业,进入教育、医疗、政务、金融等场景。越来越多人感受到AI带来的便利。从无人驾驶、机器人到脑机接口,大众对AI应用产生更多期待。

毕马威中国首席技术官及创新主管合伙人刘建刚表示,人工智能已经不是一个未来概念,对很多企业而言,利用人工智能加速决策、创新产品、优化运营,已不是选择题,而是必答题。

数据显示,2024年我国完成备案并上线提供服务的生成式人工智能大模型接近200个,注册用户超6亿;工业机器人装机量占全球过半。

今年政府工作报告提出,持续推进“人工智能+”行动。国务院研究室副主任陈昌盛表示,今年将开展新技术新产品新场景大规模应用示范行动,在确保安全前提下,加快人工智能在低空经济、教育培训、医疗健康等多场景应用。

中兴通讯股份有限公司董事长李自学表示,当前人工智能通用型应用加速普及,但在行业领域还存在不好用、不易用、不会用等问题。应在通用基础大模型上,结合不同行业应用场景需求,加强专属领域定制与深度训练,让大模型成为懂行业的“能工巧匠”,促进人工智能与产业真正结合。

有观点认为,未来可能出现自然人、机器人、数字人并存交互的社会形态。《博鳌亚洲论坛亚洲经济前景及一体化进程2025年度报告》指出,随着人工智能技术的广泛应用,部分行业就业受到严重影响。

“替代部分工作岗位是技术变革中的必然过程,有些岗位受影响,也会有新岗位出现。”vivo高级副总裁、首席技术官施玉坚说,每个人都要通过学习不断提升综合素质,适应AI带来的变化。

多位与会嘉宾表示,除了加强人工智能知识技能教育培训,还应考虑制定新的社会保障政策,加强对受冲击劳动者群体的兜底保障。

加固“安全护栏”

在博鳌亚洲论坛多场人工智能相关活动上,与会嘉宾普遍认为,随着AI能力越来越强,AI不可控、被滥用等风险会越来越高,在推广应用中要加固“安全护栏”。

社交平台上,不少人吐槽AI一本正经地胡说八道。清华大学公共管理学院院长朱旭峰表示,这是AI幻觉,指大模型有时混淆事实和虚构,在看似是事实的句子中插入错误细节。AI胡编乱造可能产生误导,甚至会引发严重后果,特别是在医疗、金融、法律等对信息真实性和准确性要求极高的领域。

“技术暂时不成熟,我们不能因噎废食。”朱旭峰说,要加大对技术的开发和使用,推动AI幻觉等问题随技术不断进步逐步解决。

AI应用过程中需要使用大量数据,数据安全与隐私保护备受公众关注。多位与会嘉宾表示,现实中,具体应用场景下哪些数据必须收集,敏感信息数据如何避免泄露等,需要进一步明晰相关制度规则。

施玉坚认为,随着AI技术的普及,企业对数据的依赖程度日益加深,如果不采取相应的安全措施,就可能造成数据泄露的风险。“数据要进行脱敏或加密处理,增强传输和存储过程中的安全性。”

AI技术的滥用误用是显著风险之一。有统计显示,2024年全球发生的AI风险事件,超过30%与利用AI进行深度伪造相关。深度伪造名人形象、声音进行虚假宣传甚至诈骗不时见诸报端。

多位嘉宾表示,AI深度伪造现象层出不穷,主要在于造假成本低,追查、执法成本大,应针对AI误用滥用完善相关法律法规,加大惩罚力度。

伴随AI产生的伦理道德争议也不容忽视。中国科学院自动化研究所研究员、联合国人工智能高层顾问机构专家曾毅认为,伦理安全应作为人工智能大模型发展的“基因”,如何在追求技术进步的同时坚守道德底线,是需要共同面对的重要课题。

“人工智能的发展与安全不是互相掣肘的关系。”曾毅说,最新研究表明,可以在几乎不影响人工智能大模型求解能力的同时,通过科学的方式提升其安全能力。

让治理跟上技术步伐

如何让治理跟上AI技术步伐?与会嘉宾表示,AI治理是一个全球性的复杂问题,需要达成世界共识,推动各领域相互协作。目前,主要经济体都在加速相关立法进程。

在AI治理中,中国坚持统筹发展和安全,有关部门已发布多项AI治理指导文件。其中,前不久发布的《人工智能生成合成内容标识办法》重点解决“哪些是生成的”“谁生成的”“从哪里生成的”等问题,推动由生成到传播各环节的全流程安全管理。

在芬兰前总理埃斯科·阿霍看来,政策制定者、企业家、科学家等利益相关方应聚在一起,建立共识,通过制定标准来应对AI带来的治理挑战。

“AI医生”看病出了问题算谁的责任?当人工智能系统产生不良后果,如何确定责任归属也需进一步厘清。“多位嘉宾都提到一个观点,不要让人工智能完全代替人的决策,这是人工智能应用时必须注意的问题。”曾毅说。

高质量数据是AI应用大规模落地的重要支撑,构建一个完善而成熟的数据市场也十分重要。以色列民主研究所高级研究员特希拉·施瓦茨·阿尔特舒勒认为,应根据实际情况,建立符合自身需要的数据监管方案与路径。

与会嘉宾表示,AI时代到来,要坚持应用与治理平衡、创新与监管并重、全球化与本土化协同,防止数字鸿沟变成智能鸿沟,让AI真正成为推动社会进步的普惠力量。

新华社海南博鳌3月27日电