必须防AI深度伪造:如何防?

虚假信息泛滥 风险不容忽视

文/姜浩峰

“确实让我更好看了。”意大利总理梅洛尼近日在社交媒体上转发她的反对者所发的一则帖子,其中有一张人工智能(AI)生成的她的性感照片。两天之后,欧洲议会议员和欧盟成员国就修订《人工智能法案》达成共识,同意禁止AI生成深度伪造的色情内容。

▲ AI聊天机器人“格罗克”被禁止生成深度伪造的色情图像

◀ 韩国政府成立跨部门机构应对数字性犯罪

欧盟立法明确禁止

以意大利兄弟党党首身份上台、成为意大利首位女总理以来,梅洛尼一直推行“意大利优先”、保守主义政策和严格的移民政策,被反对者频频使用AI生成的虚假内容进行攻击。

与此前一些移花接木的“手动P图”不同,此次有关梅洛尼的AI合成照片,人物整体协调,光影效果自然。哪怕在摄影或绘画领域有所造诣的人,也未必分辨得清照片中是否梅洛尼本人,除非与她特别亲密者。

梅洛尼在转发帖子的同时表示,“我可以为自己辩护,但许多其他人无法做到”。有评论认为,梅洛尼这一表态不仅针对“个人攻击”行为,更上升到对AI时代虚假信息泛滥的公共警示。此前她已在意大利推动通过相关法律,旨在打击针对女性的AI深度伪造图像。

两天之后,欧洲议会议员和欧盟成员国以修订《人工智能法案》的方式,为AI合成内容划定“红线”。据新华社报道,欧洲议会议员麦克纳马拉在法案修订后表示,这是欧盟首次通过立法明确禁止“脱衣换脸”类应用程序,“欧盟今天划定了‘红线’,AI绝不能用于羞辱、剥削或危害他人”。

也有评论认为,此次修订对已经产生的AI乱象的监管仍然偏软,如推迟实施针对高风险AI系统的监管规定,令人遗憾。据原计划,针对独立AI系统的监管规定应于2026年8月生效,针对嵌入其他产品的AI工具的监管规定应于2027年8月实施。然而目前来看,上述生效时间将分别推迟至2027年12月和2028年8月。

但无论如何,欧洲确实在AI监管方面又迈出了一小步。欧盟轮值主席国塞浦路斯的欧洲事务副部长劳纳表示,《人工智能法案》经过此轮修订后,“规则实施更顺畅统一,同时强化对儿童的保护”。

伪造色情照片成灾

“深度伪造”依赖于AI深度学习技术,可以把人脸、声音、动作等换到另一个图像上,让它看起来像是真实发生的事情,如让某人“说”他从未说过的话,或者“做”他从未做过的事。

早在欧洲通过《人工智能法案》的2024年,AI深度伪造就已出现在全球不少地方。如在韩国娱乐行业,不少女性从业者发现自己的形象被不法分子恶搞,甚至制作成色情内容,不仅在网络传播,也被用来勒索。为此,韩国娱乐产业公司与警察局签署协议,合作共同打击AI深度伪造。警方同意加快相关案件调查进程,设立举报热线,方便公众报告相关犯罪活动。相关企业宣布加强对深度伪造内容的监控和处理,尤其在色情内容方面,将专门设立响应机制,一经发现就采取行动。

事实上,AI深度伪造之害已经不囿于娱乐产业,也波及社会层面。2024年8月韩国某公益组织所发一组报告显示,社交平台Telegram的聊天群组中,只需上传熟人照片,付费后5秒钟内就能生成不雅合成图。令人震惊的是,这个聊天群组的参与人数高达22.7万。当时美国一家名为“家安英雄”的网络安全公司的调查报告显示,在互联网所有深度伪造视频中,几乎98%为色情视频;前十个专门用于深度伪造色情内容的网站上,视频总访问量超过3亿次,总观看次数超过3000万次。

而此后的情况越发不容乐观。自2025年12月25日至今年1月1日,美国企业家马斯克旗下社交媒体平台X的AI聊天机器人“格罗克”被指以深度伪造方式生成2万张图像,55%的图像中人物穿着暴露,其中81%是女性;2%的图像中人物年龄不足18岁。X平台随后不得不宣布,不再允许“格罗克”生成基于真人的深度伪造的性暴露图像。

亟待建构AI伦理

当今在AI领域发展较快的国家,美国是其中之一。在美国联邦调查局(FBI)公布的“网络犯罪报告2025”中,与深度伪造相关的犯罪远远不止虚假色情视频之泛滥。FBI下属机构2025年接获2.2万起AI犯罪案件,涉案总金额超过8.9亿美元,其中深度伪造技术结合AI制作钓鱼邮件和语音克隆,大幅提升了诈骗“逼真度”,显著提高了成功率。

对冲基金经理琼斯为此忧心忡忡地称,政府部门应该要求在AI生成的内容上强制添加水印,以区分真实内容与深度伪造内容。在美国政府3月发布全国性AI政策框架之后,琼斯参加了一场AI专家与模型开发者会议。他发现,与会者中有约80%支持对AI技术进行强制监管。而在2025年初,哪怕看到欧洲启动《人工智能法案》,与会者中也只有20%认为政府需要介入AI领域。短短一年,与会者的观点之所以出现逆转,很大程度上是因为大家意识到,AI在个人安全、隐私及国家安全层面存在潜在的长期风险。

在中国,AI深度伪造问题也已经引起重视。上海社会科学院上海国际经济交流中心专家郭进分析:“我国对滥用AI的现象一贯保持严格监管的态度,也积极采取各种措施严厉打击网络诈骗犯罪。利用人工智能进行网络诈骗是新情况、新问题,需要世界各国通力合作,共同打击利用AI换脸、拟声、智能黑客等网络犯罪行为,严格禁止AI武器化。”

郭进指出,AI深度伪造有违人类社会“真实性”的基本伦理,用深度伪造内容从事色情、诈骗,不仅有违公序良俗,而且违反法律,涉及犯罪。尽管各国社会文化和政治制度迥异,对AI伦理的理解千差万别,其中还掺杂许多意识形态的内容和要求,但总体上都认同“智能向善”的基本原则。“各国法律和制度都明令禁止AI干预政治,比如规定AI不得煽动暴力、不得鼓动颠覆政权等,这是各国为AI划定的伦理底线和法律红线。从维护公民人格权、隐私权、名誉权、肖像权等权利的角度出发,换脸、拟声、算法歧视等都属于应被严格监管的行为。”

郭进认为,欧盟对AI向来秉持“严格监管”的立场,对AI治理和规范的诸多理念得到关注。欧盟2019年在《可信人工智能伦理准则》中提出的“人类自主权与监督、技术稳健、隐私保护、透明度、公平性”等准则,以及2020年《人工智能白皮书》中提出的对AI分级监管的思路,至今仍值得各国借鉴。相信在不久的将来,全球对AI的监管将从伦理倡导走向硬性约束,并跃升为制度规范。

▼ 欧盟委员布雷东在欧盟《人工智能法案》通过后发表讲话

◀ 黑客用深度伪造技术“换脸” 本版图片 IC