人工智能“深度伪造”技术风险刑法规制的向度与限度*

姜　瀛

摘　要　“深度伪造”是一种以人工智能深度学习为基础的非真实音视频合成技术。基于“生成对抗网络”之技术内核，“深度伪造”技术所合成的非真实音视频作品具有高逼真度，若被不法利用确实可能引发严重危害后果。但应当看到，“深度伪造”的技术风险在于增加了虚假信息对国家安全、社会秩序或公民权利等传统法益的侵害程度，并未侵犯到新的法益类型。因此，基于法益的立法批判机能，刑法规制“深度伪造”技术的向度应当是在司法层面上降低入罪门槛与提升刑罚量，而并非是在立法层面上盲目地犯罪化。主张以“盗窃个人生物识别信息为坐标将身份盗窃行为入罪化进而规制‘深度伪造’技术”的立法论观点，脱离了立法事实而制造出法益假象，不具有正当性。新兴技术治理过程中，我们要正视刑法的功能限度，刑事高压政策可能会扼杀创新，危及到人工智能技术的发展前景。立足于开放性治理模式并建构以“标识义务”为中心的制度体系，将是“深度伪造”技术法律规制的理性选择。

关键词　人工智能；深度伪造；犯罪化；司法应对；开放治理

中图分类号　D914　　文献标识码　A　　文章编号　1001－8263(2021)09－0101－09

DOI：10.15937/j.cnki.issn1001－8263.2021.09.012

作者简介　姜瀛，大连海事大学法学院副教授、博士　辽宁大连 116023

为抢抓人工智能发展的重大战略机遇，构筑我国人工智能的先发优势，2017 年 7 月 8 日，国务院发布《新一代人工智能发展规划》。借助大数据和算法的高效结合与实践转化，人工智能发展规划逐步落到实处，对经济发展与社会生活的多个方面产生深远影响。在充分肯定人工智能技术创造经济效益与社会价值的同时，人工智能所蕴含的技术风险也引发社会的广泛关注。2017 年 12 月，有用户在美国红迪网(Reddit)论坛上传了一段技术合成的色情视频①，好莱坞影星成为色情视频“主角”，真实程度让人瞠目结舌。无独有偶，2019 年 8 月，一款名为“ZAO”的换脸软件在我国上线，下载量火爆，但很快就被下架。②近期，一款名为“Avatarify”的换脸短视频编辑软件(即“蚂蚁呀嘿”特效制作软件)上线后受到热捧，但此后很快就在中国区下架。③事实上，不论是美国的色情视频，还是国内的换脸软件，其中均运用了人工智能“深度伪造”(Deepfake)技术。应当看到，“深度伪造”技术绝非是一种单纯的娱乐工具，其所具有的技术风险已经显现出来。

直观来看，“深度伪造”是依托于大数据与人工智能深度学习所形成的智能音视频处理技术，其在艺术(历史图片修复与音视频处理)、教育以及自主学习等领域具有广阔的应用前景。④随着深度学习与生物特征数据识别等技术的不断成熟，“深度伪造”不再远离日常生活；技术上的“低

* 本文是国家社科基金青年项目“刑法立法模式与修改方式研究”(19CFX038)、中央高校基本科研业务费专项资助项目“网络数据爬取行为刑法规制问题研究”(3132021287)的阶段性成果。

门槛”使得“深度伪造”很容易被普通民众所驾驭，面临着被不法利用的风险。因此，从刑法层面对“深度伪造”这一新生事物展开思考具有其合理性。一方面，由于“深度伪造”技术的本质是制造“假象”，其所合成作品的“逼真”效果可以使虚假信息对个人名誉、商业信誉或社会秩序等固有法益造成更大的危害；当行为人利用“深度伪造”技术所实施的不法行为符合某一犯罪的构成要件，如诽谤罪、损害商业信誉罪、寻衅滋事罪、编造、故意传播虚假信息罪、传播淫秽物品牟利罪以及诈骗罪，等等，⑤便可能被以相关罪名追究刑事责任。在现有的刑法规范框架下，基于“解释论”路径，不法利用“深度伪造”技术的行为可被直接适用刑法相关罪名予以规制。另一方面，在明确不法利用“深度伪造”技术的行为存在刑事违法性的同时，“深度伪造”本身的技术风险是否已达到刑法专门规制的必要程度，也即是否有必要在立法上针对“深度伪造”技术本身作出专门回应——立法上的犯罪化，则值得进一步探讨。

近期，“从刑法立法层面对‘深度伪造’技术作出专门回应”逐步成为一种有力声音。有学者认为，“深度伪造技术滥用的根本原因在于个人生物识别信息的滥用，而我国刑法忽略了深度伪造法益侵害的独立性以及个人生物识别信息保护的特殊需求。深度伪造技术滥用规范的本质是身份盗窃行为，有必要在刑法中引入身份盗窃”。⑥另有学者指出，“刑法规制‘深度伪造’技术的正当性源于对该技术的自主性、便捷性、逼真性等特征及其隐患的审视。但合理控制刑法介入的深度，以发挥刑法的预防机能为目标。通过增强对于个人生物识别信息的刑法保护，从前端防范‘深度伪造’技术被滥用，应增设身份冒用罪”。⑦

应当看到，人工智能对人类生活和社会发展的影响可能超越历史上任何一个时代，其中引申出的新兴科技治理中的法治理性问题值得关注，而“深度伪造”技术则可以成为思考这一问题的良好样本。对于“深度伪造”这一可能被不法利用的人工智能技术，是在刑法立法上专门确立起一种高压手段，还是说选择一种相对缓和的法律治理对策，探讨刑法规制“深度伪造”技术的理性“向度”，并明确刑法在面对新兴技术时的局限性，也即“限度”，具有现实意义。

一、“深度伪造”的技术解构与规制考察

（一）“深度伪造”的技术内核与特征解析

“深度伪造”，顾名思义，即深度学习（Deep - learning）与伪造（Fake）的结合，其中的核心技术是“生成对抗网络”（Generative Adversarial Network，简称 GAN）。⑧与传统深度学习技术单链条相比，GAN 引入了“对抗”机制，由两组神经网络共同进行，其中一组神经网络的算法定位为“生成器”，它负责基于“源数据”创建目标图像模型，从而生成伪造的图像；另一组神经网络的定位为“鉴别器”，它以真实目标图像为标准对“生成器”的合成作品进行检验。⑨相较于传统的深度学习技术，“生成对抗网络”类似于存在产品检验标准的自动生产线。“生成器”合成的非真实作品被送往“鉴别器”判断其伪造的逼真度；若逼真度低于特定标准则会被退回“生成器”继续修改，经过无数次地退回与再修正之后，最终会合成出“高逼真”的非真实音视频作品。简言之，GAN 就是利用“生成器网络”与“鉴别器网络”的对抗机制，使合成的非真实音视频作品更加贴近真正版本。⑩

具体来看，利用“深度伪造”技术合成非真实音视频的流程包括如下环节。⑪首先，行为人需获得目标人物的照片或音视频等源数据，作为深度学习的训练数据。其次，算法会将源数据解码，学习目标人物的面部特征、身形或语调。再次，以目标任务数据与算法为基础，GAN“生成器”自动生成目标人物的照片或音视频；当然，这一阶段所合成的作品并不是最终的“成品”。复次，“鉴别器”将会开始工作，将合成作品与真实的“原件”进行数据对比，后提出修正意见并将其返回“生成器”，直到“鉴别器”对其判断为真——符合相似度的正负误差范围。最后，当合成作品足以达到“原件”标准时，算法将会自动生成以目标人物为对象的非真实音视频。

为充分理解“深度伪造”这一新生事物，我们可以从技术特征与应用特征两个层面来综合把握。首先，相较于传统的图像或视频修改技术

(如 Photoshop),“深度伪造”实现了技术上的飞跃,呈现如下技术特征。一是,自主生成。“无监督学习”模式下所实现的机器自主学习使得“深度伪造”技术合成非真实音视频的核心过程不需要人的参与⑫,使用者只需提供重组的目标人物源数据即可。二是,自我进化。每次合成非真实音视频的过程都伴随着自主学习,随着海量数据的填充,“深度伪造”在技术上完成了算法优化。⑬其次,基于上述技术特征,“深度伪造”在被广泛应用过程中又体现出如下应用特征。一是,操作简便。“半监督”或“全监督”模式下的深度,使用者需要自己手动标注训练数据,技术性要求较高;而“深度伪造”的无监督式学习大大降低了技术门槛,操作的简便使得越来越多的非专业人士也可以熟练运用“深度伪造”技术。二是,鉴别困难。在“深度伪造”技术出现以前,一般图像处理技术对于光线、阴影、语调等的处理无法达到“逼真”程度,通过技术容易鉴别。⑭但随着“生成对抗网络”技术的应用,只要拥有足够的源数据以及足够的训练时间,合成作品足以达到“以假乱真”的程度。

(二)我国规制深度伪造技术的现有规范

自 2019 年以来,网信办、国家广播电视总局(以下简称“广电总局”)相继出台一系列部门规章用以规范“深度伪造”技术利用行为。与此同时,2020 年 5 月十三届全国人大三次会议表决通过了《中华人民共和国民法典》(以下简称《民法典》),其中人格权编对于保护公民肖像权所作出的规定也充分关注到“深度伪造”这一技术前沿问题。目前,我国立法已经对于“深度伪造”技术利用行为作出具体规定(详见表 1),呈现如下特征。

首先,围绕不当利用“深度伪造”的行为,我国确立了行政立法与民事立法相结合的立法构造。近年来,网信办、广电总局等相关主管部门相继制定《网络音视频信息服务管理规定》(2019 年 11 月 18 日颁布,以下简称《音视频规定》)《网络信息内容生态治理规定》(2019 年 12 月 15 日颁布,以下简称《内容生态规定》)等部门规章,明确“深度伪造”不当利用的行政违法类型与制裁措施。与此同时,《民法典》第 1019 条规定,“任何组织或个人不得利用信息技术手段伪造等方式侵害他人的肖象权”。事实上,“利用信息技术手段伪造等方式”所针对的就是“深度伪造”技术,可以说,《民法典》的这一规定极具时代特征。

其次,《音视频规定》确立了“深度伪造”技术规制的专门化措施,包括标识义务、技术保障义务、停止传输义务以及辟谣机制等多个方面。具体来看,《音视频规定》第 11 条第一款规定,“网络音视频信息服务提供者和网络音视频信息服务使用者利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播非真实音视频信息的,应当以显著方式予以标识”,即“网络音视频信息服务提供者”与“网络音视频信息服务使用者”均负有对“深度伪造”音视频作品的标识义务;第 12 条规定,“网络音视频信息服务提供者应当部署应用违法违规音视频以及非真实音视频鉴别技术”,即网络音视频信息服务提供者负有“部署深度伪造鉴别技术”的义务,并在发现后停止传输信息;第 13 条规定,“网络音视频信息服务提供者应当建立健全辟谣机制”,即“应当建立辟谣机制”,在发现虚假信息后及时辟谣。可以看到,上述规定明确了我国合理使用“深度伪造”技术的边界以及规制不当使用“深度伪造”技术的专门措施。

最后,对于新闻类信息,我国作出禁止性规定。《音视频规定》第 12 条第二款规定,“网络音视频信息服务提供者和网络音视频信息服务使用者不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息”。在我国,新闻类信息具有高度的权威性,一旦行为人利用“深度伪造”技术非法合成新闻作品,则会让虚假信息以高度可信的方式呈现给社会大众,极大地冲击新闻媒体与政府机关的公信力。正因如此,我国明确禁止在新闻信息领域使用“深度伪造”技术,维护新闻的真实性,保障民众对于社会真相的知情权。

总体而言,针对“深度伪造”这一新生事物,我国以行政立法与民事立法相结合的方式作出积极回应;在肯定“深度伪造”技术具有合法性的基础上,确立了以“标识义务”为核心措施的开放性治理路径。

表 1 “深度伪造”相关法律法规

规范名称	发布时间	性质	核心内容
《民法典》第 1019 条	2020 年 5 月 28 日	法律	任何组织或个人不得利用信息技术手段伪造等方式侵害他人的肖象权。
《网络音视频信息服务管理规定》第 11 条	2019 年 11 月 18 日	部门规章	网络音视频信息服务提供者和网络音视频信息服务使用者利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播非真实音视频信息的，应当以显著方式予以标识。 网络音视频信息服务提供者和网络音视频信服务使用者不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息。转载音视频新闻信息的，应当依法转载国家规定范围内的单位发布的音视频新闻信息。
《网络信息内容生态治理规定》第 23 条	2019 年 12 月 15 日	部门规章	网络信息内容服务使用者和网络信息内容生产者、网络信息内容服务平台不得利用深度学习、虚拟现实等新技术新应用从事法律、行政法规禁止的活动。
《数据安全管理办法（征求意见稿）》第 24 条	2019 年 5 月 28 日	部门规章	网络运营者自动合成新闻、博文、帖子、评论等信息，应以明显方式标明‘合成’字样；不得以谋取利益或损害他人利益为目的自动合成信息。
《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》第 3 条第 2 款	2018 年 11 月 15 日	部门规章	互联网信息服务提供者使用新技术新应用，使信息服务的功能属性、技术实现方式、基础资源配置等发生重大变更，导致舆论属性或者社会动员能力发生重大变化的应自行进行安全评估，并对结果负责。

二、“深度伪造”技术的犯罪风险分析

“深度伪造”技术逐步为普通民众所掌握，此后便引发一系列事件（详见表 2），涉及到“政治活动、色情、诈骗以及娱乐软件”等不同类型。“深度伪造”所具有的技术风险开始为法学界所关注。有学者指出“‘深度伪造’是人工智能技术进步的产物之一，它运用的‘生成对抗网络’技术被用户快速普及，对其滥用则会威胁国家安全、个人和企业的合法权益，它严重影响了信息安全”。⑮另有学者认为，“深度伪造可能用于色情视频、虚假新闻以及虚假广告等方面，进而引发社会风险，危及个人合法权益、企业商业信誉以及社会公共安全”。⑯还有学者强调“深度伪造技术滥用不仅将侵犯公民的个人权利，破坏社会稳定与国家安全，甚至还可能消解社会共同的信任”。⑰

不可否认，“深度伪造”技术的不当利用确实存在着巨大风险，这实际上也是技术创新所面临的共性问题。学界的已有研究明确了“深度伪造”的风险领域，从国家、社会与个人的不同维度解释出其所侵犯的法益类型。不过，若仔细甄别，我们不难发现，“深度伪造”技术实际上只是增加了虚假信息对固有法益的侵害程度，并没有侵犯新的法益类型。因此，前述基于“国家社会与个人”的类型化研究视角，并未能从本质上揭示出“深度伪造”所蕴含的技术风险。应当看到，“深度伪造”的技术风险在于“量”的增强，而非“质”的改变，对于“深度伪造”技术风险的法律对策分析应当以此为坐标来展开。

表 2 “深度伪造”典型事件

时间	内容	地点	性质
2017 年	黑客攻击账号发布卡塔尔元首关于伊朗和伊斯兰教的虚假讲话	中东地区	国家安全
2017 年	加尔盖多换脸色情视频	美国	个人名誉
2018 年	Lyrebird 的公司深度合成语音	加拿大	娱乐性
2018 年	印度记者 Rana Ayyub 遭到色情报复	印度	个人名誉
2018 年	奥巴马攻击特朗普是个“十足的笨蛋”	美国	国家安全
2018 年	加蓬总统邦戈中风后发布伪造的新年致辞	加蓬	国家安全
2019 年	杨幂换脸《射雕英雄传》	国内	娱乐
2019 年	ZAO 软件	国内	娱乐
2019 年	特朗普于 Facebook 发布佩洛西如同喝醉的演讲视频	美国	政治
2019 年	Deepnude 软件一件脱衣	美国	色情
2019 年	以色列一家公司合成扎克伯格关于脸书技术垄断的虚假视频	美国	商业信誉
2019 年	犯罪人利用深度伪造模仿某能源公司 CEO 声音实施诈骗	德国	诈骗
2019 年	美国 Kneron 公司使用深度伪造技术，欺骗了支付宝和微信的支付程序，并顺利通过机场等自助终端的检验	美国	诈骗
2020 年	特朗普于推特发布佩洛西撕毁国会讲稿的虚假视频	美国	政治
2021 年	“Avatarify”换脸短视频编辑软件	国内	娱乐

第一，“深度伪造”技术所产生的“逼真”效果，将会使虚假信息对于国家安全、社会秩序或公

民权利的危害进一步放大，并危及到互联网的信任基础。应当看到，利用“深度伪造”技术合成的非真实音视频“具有高逼真度与难辨别性，一旦合成作品通过互联网广泛传播，其将打破人们“眼见为实”的固有认知，使得虚假信息对于国家安全、社会秩序以及公民权利等具体法益的侵害程度进一步加剧，模糊网络空间中“真”与“假”之间的界限，造成整个网络空间的信任危机。⑱

第二，不当利用“深度伪造”技术的行为可以实现法益侵害复合化的结果。在“深度伪造”技术普及之前，人们很难利用计算机手段将目标人物替换到色情视频中；单纯的传播淫秽信息的行为，仅可构成我国《刑法》第363条、第364条所规定的“传播淫秽物品牟利罪”与“传播淫秽物品罪”。而利用“深度伪造”技术以特定目标人物为对象所生成的色情视频却可能侵害数个法益，也即表现出侵害法益复合化的效果。具体而言，在行为人通过“深度伪造”技术所合成色情视频的行为可能符合“传播淫秽物品”的构成要件，同时还虚构被害人“色情经历”，其行为本质是捏造事实，侵犯到社会对人的积极评价与个人对自我价值的认识⑲，可能同时构成诽谤罪。

第三，“深度伪造”技术合成非真实音视频作品识别难度高，增加了监控负担。对于利用“深度伪造”技术所合成的音视频，没有经过专业训练以及相关专业技术辅助的普通民众难以准确区分。有报道指出，瑞士科学家尝试用最前沿的人脸识别系统去识别“换脸视频”，结果错误率高达95%；德国和意大利科学家的联合研究小组测试了1000段“换脸术”视频后发现，普通人必须通过特殊训练，才有可能鉴别真伪。⑳目前，由于各大网络平台尚未开发出有效的“深度伪造”技术鉴别措施，而且“深度伪造”技术又在进化发展，因此，鉴别技术的相对滞后性难以克服。可以说，鉴别难度与监控负担将在一段时间内成为“深度伪造”技术法律规制的挑战。

综上，“深度伪造”技术所蕴含技术风险在于对虚假信息造成社会危害的“加成”效果，其不当利用可能会加剧法益侵害的程度或者使法益侵害复合化，但并未侵犯到新的法益类型。

三、刑法规制“深度伪造”技术的向度展开

（一）刑法立法层面专门回应“深度伪造”的理论误区

目前，国内主张以刑法立法专门规制“深度伪造”技术的观点——立法犯罪化路径，其核心论据有两个方面，其一是“深度伪造”的技术风险很可能演化为“身份盗窃”之现实危害，其侵犯到个人生物识别信息这一有待刑法保护的法益㉑，其二是以美国为代表的域外立法已经专门将不当利用“深度伪造”技术的行为予以犯罪化。㉒对于上述观点以及论据，值得进一步探讨，目前，我们难以证实不当利用“深度伪造”技术的行为已经侵犯到某种特定的、有待在刑法中确立的新法益，盲目采取立法犯罪化的路径并不具有正当性。

首先，“深度伪造”技术的不当使用虽可能侵犯到个人生物识别信息——如某换脸视频便是利用了他人面部信息，但此种行为未必一定构成“身份盗窃”，将盗窃个人生物识别信息作为刑法规制“深度伪造”的立法事实并不妥当。目前，研究者预期中的不法利用“深度伪造”技术而严重侵犯个人生物识别信息的典型例证，即伪造他人面部生物信息而突破“人脸识别”措施——构成身份盗窃，最终利用他人身份实施不法行为（如侵财犯罪）。但从实践情况来看，行为人多是利用后台程序便可以突破“人脸识别”措施，没有必要利用“深度伪造”技术将犯罪过程复杂化。同时，单纯以明星为目标人物而进行的换脸视频，仅仅是一种娱乐行为，显然无法达到身份盗窃的法益侵害程度，由民事法律规范来调整即可。因此，以“深度伪造”可能侵犯个人生物识别信息并成立身份盗窃的理论假设，目前尚不能被证明是明确的立法事实。此外，将身份盗窃行为犯罪化所要保护的并不仅仅是个人法益，更在于个人法益之上的以社会信任为基础的社会秩序。因此，即使行为人以“深度伪造”突破“人脸识别”措施并非法取得他人财物，该不法行为并未骗取社会的信任——并未向社会大众展现其冒充的身份，其行为本质只不过是欺骗“机器”取得财物而已。正因如此，2020年12月26日，十三届全国人大

常委会第二十四次会议通过的《中华人民共和国刑法修正案(十一)》第 32 条[23]所增设的“冒名顶替罪”,仅将盗用、冒用身份行为明确限定在“顶替他人取得的高等学历教育入学资格、公务员录用资格、就业安置待遇”等具有社会意义的升学就业领域,并将该罪置于刑法分则第六章“妨害社会管理秩序罪”一章,其本意旨在保护社会对于身份的信任,而非单纯的个人法益。

其次,域外现实状况与我国存在明显区别,我国并不具有以刑法专门规制“深度伪造”技术的现实必要性。从世界范围来看,美国是针对“深度伪造”进行专门立法的代表性国家,一些州针对不当利用“深度伪造”的行为迅速做出反应,通过颁布法案或修法实现犯罪化。具体来看,美国针对“深度伪造”的州立法集中在“色情视频”与“涉及政治的选举活动”两个领域。在 2019 年,弗吉尼亚州通过修改《复仇情色法案》,将利用“深度伪造”技术合成他人色情视频并予以传播的行为犯罪化;得克萨斯州颁布了《关于伪造欺诈视频影响选举结果的刑事犯罪法案》,将利用“深度伪造”技术合成候选人虚假视频并予以传播的行为犯罪化;加利福尼亚州也颁布了《第 730 号议会法案》,将利用“深度伪造”技术合成视频干扰选举行为犯罪化。[24]值得注意的是,美国针对“色情视频”整体上采取了合法化策略,“色情视频”具有正当地位,但需要实施分级管理。[25]正是由于“色情视频”本身不具有违法行为,针对利用“深度伪造”技术合成非真实“色情视频”所产生的社会危害,才有专门犯罪化的必要性。反观我国,传播淫秽物品行为具有明确的刑事违法性,因而不论是否使用“深度伪造”技术,均构成犯罪。因此,对于“深度伪造”技术合成他人色情视频,我国无需进行专门刑事立法予以犯罪化。同时,由于我国针对涉及政治活动、政府形象以及其他涉及公共秩序的信息采取事前管控模式,无论是网络平台服务提供者还是执法机关都采取严格的审查政策,利用“深度伪造”合成涉及政治活动、政府形象以及其他涉及公共秩序的非真实音视频并予以传播的行为,在实践中难以实施。在相关行为已经被全面规制的前提下,简单借鉴域外立法主张对“深度伪造”技术作出刑法立法回应,脱离了本土话语体系,不具有正当性。

(二)刑法司法层面回应“深度伪造”的路径展开

如前所述,“深度伪造”技术的不当利用可能会加剧法益侵害的程度或者使法益侵害复合化,但并未侵犯到新的法益类型。[26]因此,刑法对“深度伪造”技术的理性回应应立足于司法层面,也即通过适用现行刑法条文来实现对“深度伪造”的有效应对,而并非在立法上盲目的犯罪化。具体而言,这种司法应对路径应以社会危害程度为基础,包括降低犯罪门槛与增加刑罚量两个不同面向。

首先,针对利用“深度伪造”技术实施的犯罪行为,刑法适用过程中应降低相关犯罪的入罪门槛。应当看到,利用“深度伪造”技术合成的非真实音视频具有高逼真度,由此提升了虚假信息的可信性与辟谣难度,进而使编造、传播虚假信息行为的危害性显著增强。而根据最高人民法院、最高人民检察院于 2013 年 9 月 6 日联合颁布的《关于办理利用信息网络实施诽谤等刑事案件适用法律若干问题的解释》(以下简称《网络诽谤解释》)第 2 条第一款规定,一般情况下,只有同一诽谤信息实际被点击、浏览次数达到五千次以上,或者被转发次数达到五百次以上的,即该视频被浏览五千次以上或者转发五百次以上,才能符合诽谤罪“情节严重”的入罪条件。[27]不过,考虑到利用“深度伪造”对于网络虚假信息的“催化剂”效应,刑法适用过程中可以跳出浏览或者转发次数的限制,直接将不法利用“深度伪造”技术认定为《网络诽谤解释》第 2 条第(四)项所规定的“其他严重的情形”。易言之,在利用“深度伪造”技术实施编造、传播虚假信息行为的过程中,“深度伪造”技术本身可以解释为“其他严重的情形”之入罪标准,而无须机械地以虚假信息的浏览或者转发次数作为入罪标准。

其次,针对利用“深度伪造”技术所实施的犯罪行为,我们在刑罚裁量时,不仅仅要考虑传播范围(浏览或者转发次数),更应当将其逼真程度(也即鉴别难度或澄清成本)作为危害程度的实

质认定标准。不可否认，重刑主义思想在我国长期存在；面对高发的新型网络犯罪，司法实务部门不能一味地倾向从重处罚。然而，若是新型犯罪引发的严重社会危害未能在既有司法解释的量刑情节有所体现，那么，裁判者便应当将之作为一种酌定从重情节予以规制。而利用“深度伪造”技术合成非真实音视频的高逼真度显然提升了虚假信息的危害性，裁判者对此不能视而不见。本文认为，基于罪责刑相适应原则[28]，对于利用“深度伪造”技术合成非真实音视频所形成的虚假信息，在对行为人量刑时应确立起“传播范围”+“逼真度”的复合式评价标准，“传播范围”属于客观标准，“逼真度”则是基于裁判者的主观判断，二者相结合确立酌情从重处罚的裁量标准。

四、刑法规制“深度伪造”技术的限度把控

应当承认，主张在司法层面确立刑法回应“深度伪造”技术的基本向度——降低入罪门槛与提升刑罚量，这也是基于刑法作为事后性的理性定位。客观而言，面对新兴技术，刑法具有其局限性，刑法规制的功能不应被过分强调。

首先，从新兴技术规制的法治理性层面来看，刑法在回应“深度伪造”问题时理应有所“收敛”。“深度伪造”技术是人工智能的社会化成果，通过大数据运用与算法优化实现了高逼真模拟他人动作与声音的效果——合成的非真实音视频，其被应用于艺术（历史图片修复与音视频处理）、教育以及自主学习等多个领域，并不是专门为犯罪而生的工具。当新兴技术可用于合法的、不受争议的用途时，即使其具有被不当利用构成侵权或犯罪的风险，我们也不能以刑法手段强加束缚，这是鼓励创新的必然要求。事实上，犯罪是与社会发展相伴相生的事物，科技进步必然会带来新兴犯罪形态——互联网就是最好的例证，我们应辩证看待新兴技术的犯罪风险。[29]自互联网出现以来，犯罪向网络延伸、在网络异化的态势十分明显，网络犯罪高发已经成为不争事实；即便如此，我们仍能逐步理性认识网络犯罪这种客观的、必然的社会现象，而不是“因噎废食”，更不希望通过付出制约互联网行业发展的代价来实现消灭网络犯罪的效果。应当看到，过于严苛的刑事高压政策很可能会扼杀创新，使我国在与欧美等发达国家的科技竞争中处于劣势地位，影响到我国在人工智能领域的技术话语权。

其次，考虑到刑法的保障法地位及其所应具有的谦抑性，刑法在回应“深度伪造”问题时也必然不能过于积极。对于“深度伪造”这一新兴事物，如果我们事先未能对其技术特征作出准确认知，就贸然选择刑法跟进，这不仅不利于技术创新与社会进步，更是“工具主义”刑法观的体现，必将妨碍刑法机能的优化与刑法理论的发展。与其他部门法相比，刑法作为一种事后法、保障法，这种谦抑性理念并不仅仅是一种口号，其在实质上所要考虑的是，社会治理中的特定领域都具有其自身规律，与之相应的便是专门化的法律法规体系。当我们尚未能准确认知新生事物的本质时，刑法不能成为克服恐慌的手段；即使新生事物具有某种犯罪风险，也不能在不了解对象问题时盲目主张刑事立法。更何况，针对“深度伪造”技术，我国采取了以标识义务为中心的开放性治理模式，在肯定“深度伪造”技术合理存在与规范使用的前提下推动技术应用与发展。因此，只要认同“深度伪造”这类新兴技术存在的必要性，我们便应当首先在行政法律与民事法律领域寻求科学的治理方式，切不可因为其他部门法尚未制定、对新生事物难以应对时，便随意主张采取刑事高压政策。简言之，脱离了专门性立法，简单的刑事高压政策无法从根本上解决技术本身的风险隐患。

五、结语

针对“深度伪造”这一新生事物所隐藏的技术风险，我们可以确立不同的刑法规制立场，这取决于我们对“深度伪造”技术本身的定位。一方面，如果我们认为“深度伪造”是一种有害技术，那么就应当在刑法立法上确立其违法性，可直接全面禁止或仅对特定的授权者开放使用，由此便可以达到规制效果。基于这种在刑法立法上所直接确立的违法模式，我们同时应推定民众具有违法认知与守法意愿，网络平台上传播的音视频作品通常也不会是利用“深度伪造”技术合成的违

法作品。因此,立法上也就没有必要要求网络平台特别负担对“深度伪造”技术部署鉴别的监管义务,这有助于减轻网络平台的运营负担。

另一方面,如果我们认为“深度伪造”是一种中立技术,并且是一种需要适度控制的技术,我们可以要求使用者对“深度伪造”技术合成作品予以标识,也可以将这种标识义务附加给网络平台。在这种规制模式下,未标识的音视频是否属于“深度伪造”产品,应由网络平台负责监管。刑法本身无须直接对“深度伪造”技术作出规制,但可以通过《刑法》第 286 条之一“拒不履行信息网络安全管理义务罪”㉚对网络服务提供者不履行网络监管义务的行为予以规制,进而实现对“深度伪造”技术间接规制的效果。㉛当然,若是将“深度伪造”理解为一种中立且无须控制的技术,立法上则应当恪守“通知—删除”这一基本原则,进而减轻网络服务提供者的负担并保障网络空间中的表达自由。

目前,以《音视频规定》为基础的“深度伪造”技术规制立法采取了以标识义务为中心的开放性治理对策,大体上是将“深度伪造”定位为一种需要适度控制的中立技术。因此,刑法对于“深度伪造”技术应当相应地采取间接规制——通过对网络服务提供者的刑法规制来间接实现对“深度伪造”技术的规制效果。将来,针对“深度伪造”技术,法律规制的重点应当放在标识义务的实现上,也即围绕着“履行标识义务的具体方式、履行标识义务的监管模式以及违反标识义务的补救措施与法律后果”等内容寻求科学的制度建构。只有将标识义务落到实处,才能克服刑法规制的局限性,在科技创新与风险规制之间实现最佳平衡点。

注:

①参见魏书音、刘玉琢《深度伪造技术正在颠覆网络空间的可信性》,《中国计算机报》2020 年 9 月 7 日。

②参见斯涵涵《别让换脸软件成为隐私泄露“黑洞”》,《中国妇女报》2019 年 9 月 2 日。

③“蚂蚁呀嘿凉了,为什么 AI 换脸都难逃‘英年早逝’的宿命”http://www.myzaker.com/article/6042e7488e9f092b5c7266cc/。最后访问时间 2021 年 3 月 18 日。

④㉔Bobby Chesney, Danielle Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, *California Law Review*, Vol. 107, pp. 1769 – 1770, pp. 1802 – 1803.

⑤参见姜瀛《网络寻衅滋事罪“口袋效应”之实证分析》,《中国人民公安大学学报》(社会科学版) 2018 年第 2 期。

⑥⑫㉑李怀胜《滥用个人生物识别信息的刑事制裁思路——以人工智能“深度伪造”为例》,《政法论坛》2020 年第 4 期。

⑦㉒李腾《“深度伪造”技术的刑法规制体系构建》,《中州学刊》2020 年第 10 期。

⑧⑪Robert Chesney, Danielle Keats Citron, 21st Century – Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security, *Maryland Law Review*, Vol. 78 , No. 4, p. 884, pp. 884 – 885.

⑨⑭B17 王禄生《论深度伪造智能技术的一体化规制》,《东方法学》2019 年第 6 期。

⑩余和谦《人工智慧之治理——以深度伪造为例》,《科技法律透析》2019 年第 8 期。

⑬Samantha Cole, People Are Using AI to Create Fake Porn of Their Friends and Classmates, MOTHERBOARD (Jan. 26, 2018), https://motherboard.vice.com/en_us/article/ev5eba/ai – fake – porn – of – friends – deepfakes.

⑮陈昌凤、徐芳依《智能时代的“深度伪造”信息及其治理方式》,《新闻与写作》2020 年第 4 期。

⑯张涛《后真相时代深度伪造的法律风险及其规制》,《电子政务》2020 年第 4 期。

⑱孟雪《深度伪造技术给网络可信身份管理带来的挑战与对策》,《网络空间安全》2020 年第 6 期。

⑲周光权《刑法各论》(第 3 版),中国人民大学出版社 2016 年版,第 65 页。

⑳宣晶《视频“换脸术”走近大众,引爆亿级流量后为何令人担忧》,《文汇报》2019 年 2 月 27 日。

㉓该条规定“盗用、冒用他人身份,顶替他人取得的高等学历教育入学资格、公务员录用资格、就业安置待遇的,处三年以下有期徒刑、拘役或者管制,并处罚金”。

㉕1969 年美国最高法庭对“斯坦利诉佐治亚州”(Stanley v. Georgia)一案的判决,以大法官马歇尔为首的美国最高法庭法官根据美国宪法第一和第十四修正案,确定美国人拥有和观看成人电影的权利和自由,同时也对色情视频的内容进行限制。在 1990 年“奥斯本诉俄亥俄州”(Osborne v. Ohio)一案中,美国最高法庭就把拥有儿童黄色影片定为有罪,由于美国存在电影分级制度,只要内容不涉及儿童色情,色情视频、成人电影等可以在进行分级的情况下向相应年龄段的人群播放。

㉖参见叶良芳、武鑫《法益概念的刑事政策机能之批判》,《浙江社会科学》2020 年第 4 期。

㉗《网络诽谤解释》第 2 条规定“利用信息网络诽谤他人,具有下列情形之一的,应当认定为刑法第二百四十六条第一款规

定的‘情节严重’:(一)同一诽谤信息实际被点击、浏览次数达到五千次以上,或者被转发次数达到五百次以上的;(二)造成被害人或者其近亲属精神失常、自残、自杀等严重后果的;(三)二年内曾因诽谤受过行政处罚,又诽谤他人的;(四)其他情节严重的情形。

㉘《刑法》第5条规定,“刑罚的轻重,应当与犯罪分子所犯罪行和承担的刑事责任相适应”。

㉙参见卢建平、姜瀛《论犯罪治理的理念革新》,《中南大学学报》(社会科学版)2015年第1期。

㉚该条规定,“网络服务提供者不履行法律、行政法规规定的信息网络安全管理义务,经监管部门责令采取改正措施而拒不改正,有下列情形之一的,处三年以下有期徒刑、拘役或者管制,并处或者单处罚金:(一)致使违法信息大量传播的;(二)致使用户信息泄露,造成严重后果的;(三)致使刑事案件证据灭失,情节严重的;(四)有其他严重情节的。单位犯前款罪的,对单位判处罚金,并对其直接负责的主管人员和其他直接责任人员,依照前款的规定处罚。有前两款行为,同时构成其他犯罪的,依照处罚较重的规定定罪处罚”。

㉛参见姜瀛《“以网管网”背景下网络平台的刑法境遇》,《国家检察官学院学报》2017年第5期。

〔责任编辑:木　槿〕

On the Dimension and Limitation in Criminal Regulation of AI “Deepfake”

Jiang Ying

Abstract “Deepfake” is a kind of audio and video processing technology based on the “Generative Adversarial Network” formed by the Deep-learning of big data and artificial intelligence, which can synthesize non-real audio and video works with high fidelity. Of course, “Deepfake” technology is also facing the risk of illegal use. Objectively speaking, the technical risk of “Deepfake” is that it aggravates the infringement of false information on traditional legal interests such as national security, social order or civil rights, but it does not touch the new types of legal interests. Therefore, based on the legislative critical function of legal interest, the direction of criminal law to regulate the technology of “Deepfake” should be to reduce the threshold of conviction and increase the amount of penalty at the judicial level, rather than to adopt the criminal path of blindly adding charges at the legislative level. The legislative viewpoint of adding a new crime with the coordinate of “stealing personal biometric information” so as to regulate the technology of “Deepfake” breaks away from the legislative facts and creates the illusion of legal interest, which is not legitimate. In the process of emerging technology governance, we should face up to the function limit of criminal law. Criminal high pressure policy may stifle innovation and endanger the development prospect of artificial intelligence technology. Based on the open governance mode and constructing the system centered on “identification obligation”, it will be a rational choice for the legal regulation of “Deepfake” technology.

Key words: AI; deepfake; criminalization; judicial response; open governance