AI换脸的技术风险与多元规制

林爱珺，林倩敏
（暨南大学新闻与传播学院，广东广州 510632）

摘 要：近几年，AI换脸利用人脸识别技术和深度学习算法对图片或视频中的目标人脸进行置换，在娱乐、影视行业被广泛应用，甚至被应用到新闻传播中。AI换脸不仅具有极高的欺骗性，也带来个人信息安全、肖像权的侵害及视觉失真与新闻失实等一系列社会风险。基于AI换脸的技术路径，结合《互联网信息服务深度合成管理规定》及《民法典》等相关法律法规，应当加强对深度合成技术的监管，提高对人脸生物识别信息的保护，增强平台对深度合成内容的监测并对深度合成信息内容进行显著标识。同时，要不断增强公众的数字素养和媒介使用能力，提高对深度合成内容的辨别能力，增强法制思维，维护风清气正的媒介生态环境。

关键词：AI换脸；深度伪造；肖像权；个人信息安全
中图分类号：D922.8 **文献标识码**：A **文章编号**：2096-8418（2023）01-0060-10

2019年，自“ZAO”软件应用AI换脸技术成为网络舆论焦点后，该技术也因此进入大众视野。如今，AI换脸等深度伪造技术已经覆盖了娱乐、媒体、政治等多个领域。AI换脸的虚假新闻也带来了各种新闻伦理风险，严重扰乱了舆论生态，冲击媒体公信力。2022年11月25日，国家互联网信息办公室正式公布《互联网信息服务深度合成管理规定》（以下简称《规定》），并于2023年1月10日起开始施行。《规定》对于深度合成技术在互联网中的应用作出具体说明与严格规定，回应并补充了《民法典》及《个人信息保护法》中关于利用信息技术手段伪造等方式侵害他人权利及个人信息保护的规定，对人脸生成、人脸替换、人脸操控、姿态操控等虚拟人物图像的开发与应用作出了规定，为AI换脸等深度合成技术的应用厘清底线。

一、AI换脸技术：人脸图像的深度伪造

AI换脸技术（AI face swap）是基于人脸图像识别的深度伪造技术。它源自Photoshop技术对人脸图像进行的编辑改造，广泛应用于证件照人脸美化以及人脸图像视频制作。后来，在深度学习技术的支持下，识别目标对象的面部特征而后利用算法将其“嫁接”到被模仿对象上的技术被称为人脸深度伪造技术，即AI换脸。

AI换脸技术首先建立在人脸识别技术的基础之上。人脸识别技术是识别人脸生物特征并以此进行匹配和身份识别的技术。其通过扫描人脸图像或视频，识别和侦测主要器官后形成位置信息（人脸模版），据此创建数据集并将其与待检验的人脸图像相比对，从而实现身份识别与身份验证的目的。[1]伪造人脸的深度学习算法普遍使用的是“生成性对抗网络”（Generative Adversarial Network，GAN）和卷积神经网络（Convolutional Neural Network，CNN），GAN是一种通过生成模型和判别模型互相博弈的方法来学习数据分布的生成式网络，CNN则是一类包含卷积计算且具有深度结构的前馈神经网络。[2]

基金项目：国家社科基金重大项目《人工智能时代的新闻伦理与法规》（18ZDA308）。
作者简介：林爱珺，女，教授，博士生导师；林倩敏，女，硕士研究生。

AI 换脸技术是深度伪造技术最广泛的应用。深度伪造技术（Deepfake）指基于人工智能尤其是深度学习的人像合成技术，现已发展为包括人脸图像及视频伪造、声音伪造等深度合成技术。[3] 该技术因美国“Deepfake”事件引起广泛关注。2017 年名为“Deepfake”的 Reddit 论坛用户发布了一段将斯嘉丽·约翰逊等女演员面孔嫁接在色情表演者裸体身上的假视频，用展示猎奇的淫秽表演赚流量。“Deepfake”还发布了一种机器学习算法，并声称该算法可以将名人转换成色情视频里的人物角色，引起了一阵轰动，这也意味着 AI 换脸技术算法代码的开源。此后，AI 换脸几乎成为一种无成本且简单可及的技术。在 Github 社区中 AI 换脸项目数量相当多，如 Deepface Lab、Openface Wap、Face App 等，大部分 APP 提供一键式的换脸操作，增大了技术滥用的风险。

二、AI 换脸技术侵害个人信息安全

AI 换脸技术中涉及的人脸信息包括人脸生物特征以及人脸模板，属于个人生物识别信息，具有高度敏感性。在 AI 换脸的过程中，对人脸生物识别信息的采集、存储以及使用全链条生产模式是该技术运行的底层逻辑，监管不力则存在侵害个人信息安全的风险。

（一）人脸生物识别信息具有高度敏感性

人脸生物识别信息属于生物信息，具有高度敏感性。根据《通用数据保护条例》第九条规定，生物信息包括自然人的指纹、声音、虹膜、脸相、静脉等生物信息，且均属于敏感信息，人脸生物识别信息包括人脸模版，即“脸相”，属于敏感信息。我国《个人信息保护法》规定，“个人敏感信息”指自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、金融账户、行踪轨迹以及未满 14 周岁的未成年人的个人信息等。人脸生物识别信息是对面部特征的识别信息，即生物识别信息的一种，也归属于“个人敏感信息”。

人脸生物识别信息的高度敏感性在于该类信息的唯一特性，一旦采集将很难被消除或修改。“人脸”作为人体的身体器官，具有唯一、不可替代的特性，人脸生物识别信息通过扫描人脸而得来，延续了真实人脸的唯一特性。人脸生物识别信息源于人脸模版的数据点集合，如若人脸特征本身没有发生改变，那么人脸生物识别信息也不会发生改变。人脸生物识别信息以数据集的形式存放在虚拟网络数据库中，消除与修改数据集需进入该网络节点进行删除，如若数据集已进行运输流转，那么消除数据集更是需要经由追溯流转源头或泄露源头才能查清，流程繁琐且需耗费大量人力物力。

高度敏感性也意味着存在巨大的安全风险。人脸生物识别信息能够直接链接到其他的敏感个人信息，一旦该类信息被非法盗取便可能导致其他的深层信息被挖掘与泄露，进一步造成银行卡、社交账户的非法盗用，严重威胁公民信息安全与生命财产安全。2018 年某网络诈骗团伙使用软件将公民头像照片制作成 3D 头像，并利用非法得来的公民个人身份信息注册支付宝账号，通过人脸识别认证后诈骗支付宝提供的邀请注册新用户的红包奖励，从中非法牟利 4 万元。① 不法分子将人脸照片制作成 3D 头像，利用对人脸生物识别信息的侦测原理破解支付宝的人脸识别技术，并利用非法获取的个人身份信息伪造真实个人身份实施诈骗活动。其中人脸生物识别信息的非法利用是实施诈骗活动的关键一环，利用人脸生物识别信息与个人身份信息结合伪造虚拟身份 ID，进而导致他人个人财产被盗取。随着人脸生物识别信息的应用场景日益广泛，生物信息安全风险也日益增长。

（二）AI 换脸技术的滥用增大个人生物信息安全风险

AI 换脸技术的运行过程亦是对生物识别信息（敏感信息）的处理应用过程。AI 换脸首先需要利用人脸识别技术对人脸图像进行侦测和识别，并从中提取核心特征信息，再对人脸特征进行测量得到人

① 中国裁判文书网《张富、余杭飞、史良浩等侵犯公民个人信息罪二审刑事裁定书》。

脸模板并以数据集形式储存于数据库；其次通过机器学习训练确定人脸变换矩阵，得到人脸替换转换编码；最后在原有的人脸图像上进行转化，导出“换脸”之后的伪造图像或视频。可见，识别人脸生物特征以及提取人脸模版是AI换脸技术运行的关键核心。据《人脸识别规定》第一条第三款规定，人脸识别中的人脸信息属于《民法典》第一千零三十四条中规定的生物识别信息。因此，在AI换脸技术运行过程中，对人脸生物识别信息的采集、存储以及使用全链条生产模式是该技术运行的底层逻辑，一旦发生技术操作失误或黑客入侵事件则将导致大量的个人生物识别信息泄露，存在较大的生物信息安全隐患。

而AI换脸技术的滥用更加剧了安全风险。AI换脸在当下几乎成为一种无门槛使用技术，往往作为一种娱乐手段与社交媒体音视频传播相结合，实质上是化为夺人眼球的噱头以捕获更多流量。2021年，抖音APP上发起名为性别反转挑战的话题并提供性别反转的“AI换脸”滤镜。该滤镜便是利用AI换脸技术生成对应的异性人脸图像，操作十分简单快捷，目前该滤镜的参与量已超过两千万次。用户出于娱乐休闲的目的，利用AI换脸技术制作短视频，并将其上传至网络，但对于运用AI换脸技术采集自身人脸生物识别信息的过程一无所知，更对其中生物信息采集及泄露的风险毫无察觉。

而当下我国针对AI换脸技术处理该类信息的流程仍缺少监管。目前我国有关AI换脸技术的法律规定体现在三处，分别在《民法典》《网络生态治理规定》及《网络音视频信息服务管理规定》。《网络生态治理规定》《网络音视频信息服务管理规定》这两份法律文件的规定过于宽泛，而《民法典》则主要体现在第一千零一十九条中“利用信息技术手段”侵害肖像权，这就导致其对AI换脸技术侵害个人信息安全的规制并不能发挥具有针对性的遏制作用。另外，《规定》第十四、十五条规定，对于提供具有对人脸生物识别信息编辑功能的算法模型等应当自行开展安全评估，采集人脸生物识别信息前应获得单独同意及加强储存数据集的管理。

在我国现行法律下，由AI换脸技术带来的个人生物识别信息安全威胁仍缺乏有力规制。应当加强对人脸生物识别信息的保护，对人脸生物信息的采集、存储应当采取严格的保护措施，严格限定使用场景，否则存在严重的生物信息安全隐患。

三、AI换脸侵犯肖像权

《民法典》在第一千零一十八条第二款中第一次明确规定肖像的概念与具体内涵，“肖像是通过影像、雕塑、绘画等方式在一定载体上所反映的特定自然人可以被识别的外部形象”。一般认为，肖像应当包含自然人“可识别”的外部形象，即得以在外部形象上与特定自然人之间建立对应联系。面部、身影甚至真人表情包等能够达到“可识别”程度的同样可以作为肖像。

构成肖像权侵权有两种途径：破坏他人肖像完整性，未经肖像权人同意而使用、公开他人肖像。破坏他人肖像完整性指的是采取以丑化、污损，或者利用信息技术手段伪造等方式侵害他人肖像。AI换脸技术是对人脸图像的嫁接与挪移，利用AI换脸技术伪造甚至丑化、污损，或未经他人同意使用其肖像进行AI换脸并公开都构成侵权。

（一）利用信息技术手段伪造他人肖像

《民法典》第一千零一十九条第一款规定，“任何组织或者个人不得以丑化、污损，或者利用信息技术手段伪造等方式侵害他人的肖像权”。“利用信息技术手段伪造”中的“信息技术”泛指在信息科学的指导下扩展人类信息功能的技术总和，具有一般性的技术性和区别于其他技术的信息性特征。[4]利用信息技术性的手段对他人肖像进行“再现”达到“可识别”的程度，即是“伪造”他人肖像。AI换脸技术通过扫描人脸图像提取有用的生物特征信息（如脸部轮廓、眼睛的大小、嘴巴的位置）生成人脸模版，再对人脸模版进行编码形成数据集，并依据特定的算法规则转化，实质上是将人脸生物特

征信息化并加以算法函数化转变的技术处理流程，兼具信息性与技术性。换脸后达成以假乱真、能够被识别为“同一人”的视觉效果，构成对他人肖像的“伪造”。因此，利用 AI 换脸技术将他人的肖像伪造到特定场景中的行为，属于“利用信息技术手段伪造”侵害他人肖像权。

2021 年 8 月，腾讯 QQ、微信平台上有大量用户通过私聊、群聊渠道以截图、转发形式传播含有使用 AI 换脸技术对明星刘某进行丑化、侮辱的合成视频、视频截图及含有侮辱诽谤言论的聊天记录，对当事人造成了极其恶劣的影响，严重侵害当事人的肖像权。该用户利用 AI 换脸技术将刘某的肖像嫁接到特定的淫秽场景中，即通过“信息技术手段”达到丑化、侮辱刘某的非法目的，而鉴于刘某作为明星的公众人物身份，合成的图像与视频只要足以被一般公众识别为同一人，达到可被认定为“真实”的程度，即构成“伪造”，属于“利用信息技术手段伪造”侵害他人肖像权的非法行为。

利用 AI 换脸技术对脸部图像进行嫁接、编辑，无疑破坏了肖像的完整性，是既定的侵害事实。但 AI 换脸对人脸图像的“伪造”并不等同于“丑化”“污损”他人肖像，两者的损害程度相去甚远。在 2017 年美国 Reddit 网站换脸色情视频事件中，以好莱坞当红女星为首的十几名女性遭受肖像权、名誉权的严重侵害。而在国内“肌肉金轮”系列视频中，AI 换脸技术却为当事人增加不少收益。哔哩哔哩平台某用户将某知名游戏主播的面部肖像嫁接到了以身材为卖点的国外某短视频博主的身体上，两者适配度极高，从而引发了视频的爆火，该游戏主播也受益于此知名度暴涨。利用 AI 换脸技术“伪造”他人肖像合成音视频进行传播，可能给当事人带来聚光灯式的关注与巨额的流量曝光，但利用该技术手段达成“丑化”“污损”的非法目的则直接导致当事人权益严重受损，技术使用者的主观动机与目的是其中的关键。

AI 换脸技术并非中立，而是带有“伪造”的特殊属性，但将 AI 换脸技术矮化为邪恶的技术则是典型的“技术有罪论”。技术无罪，人难脱其罪。在 AI 换脸技术开发的初始阶段，研发者们忽视了技术在发展与应用过程中存在的风险，片面考虑到技术带来的市场效益便将其投入市场，并非技术有罪，而是资本无限逐利的“原罪”。

（二）未经他人同意使用肖像进行 AI 换脸

《民法典》第一千零一十九条第二款规定，未经肖像权人同意制作、使用、公开其肖像，破坏了肖像权的专有性，具有侵害他人肖像权的违法性。在《民法典》颁行前，依据《民法通则》第一百条的规定，未经肖像权人同意而以营利为目的制作、使用、公开他人肖像的行为构成侵权行为。《民法典》颁布后不再将其作为侵害肖像权的构成要件，实质上扩大了肖像权保护范围并优先保护肖像权的精神性利益。即便没有营利目的，未经他人同意擅自利用 AI 换脸技术合成他人肖像甚至广泛传播，一般是行为人有意识的、存有某种心理动机的行为，通常不会因为不注意的心理状态而误用他人肖像，过错形态为故意，主观过错明显，客观上也造成了对他人肖像权特别是肖像权中精神利益的侵害。

未经他人同意擅自利用 AI 换脸技术嫁接他人肖像，合成“AI 换脸音视频”并散播，一般出于不同的心理动机。“AI 换脸音视频”也可以由此分为以下不同种类：娱乐类、淫秽类、政治类。娱乐传播类音视频是出于娱乐游戏化传播目的，利用 AI 换脸技术对人脸图像进行嫁接处理，最终实现在媒体上的大范围传播。淫秽类音视频是基于 AI 换脸技术将他人脸部图像嫁接至扮演色情淫秽音视频的角色身上，通常上传至色情网站并非法获利。政治类音视频是出于表达政治诉求或制造政治舆论冲突的动机，利用 AI 换脸技术将国家领袖或相关政治人物的人脸图像嫁接、处理、生成的合成音视频。

在互联网上，公众一般出于娱乐恶搞的目的利用 AI 换脸技术合成音视频，将他人的面部图像嫁接至其他人的身体上或者虚拟人物上，刻意制造反差极大或者适配度极高的视觉效果，并将合成音视频在社交媒体平台公开传播，属于“鬼畜类”二次创作。在利用 AI 换脸技术嫁接他人肖像技术的使用过程中，当事人对选择不同的面部图像应当具备一定的自主性，并非处于疏忽大意或不注意的心理状态。

哔哩哔哩平台某UP主利用AI换脸技术将明星杨某的脸叠加到影视剧中“黄蓉”一角身上并公开这段伪造视频，杨某换上了“黄蓉”的微表情后显得无比灵动，与其平时的演技形成鲜明对比，也因此引发网友的关注。因此，虽然该UP主声称制作视频并不存在营利行为，但并非出于“不注意”的心理状态，其主观动机存疑，虽无明显恶意但也难谓善意。

利用AI换脸技术将女性明星偶像的面部图像嫁接至淫秽色情视频的行为，具有强烈的丑化、侮辱等恶意，并且通常将其上传至色情网站牟利，显然构成肖像权侵权，同时还严重污损当事人名誉。对于存有政治性目标的特殊群体而言，技术是一条攫取政治利益的新途径。利用AI换脸技术将国家领导人或其他政治人物的脸部图像嫁接至其他身体，或在原有基础上单独对微表情、微动作进行转变，辅以声音轨道的智能转换合成逼真的音视频后公开，能够引发大量关注或攻击政治对手。

出于不同目的与心理动机而未经他人同意利用AI换脸技术使用他人肖像将造成不同程度的损害结果，但一般情况下都并非出于“无意识”的过失状态。对于一般大众而言，AI换脸技术更多的是一种娱乐恶搞、表达自我、休闲放松的方式，考虑到此情形下制作合成音视频的行为动机恶意不大，应当结合过错程度、受侵害人的知名度等不同方面认定侵权责任及赔偿，酌情处理。

四、视觉失真与新闻失实

真实是新闻的生命，真实的新闻内容能够帮助人们正确认识世界。真实是信任之源，真实地报道新闻能够赢得人们对媒体的信任，确立媒体合法性。但AI换脸技术的滥用带来了大量低俗、博眼球的换脸音视频的无序传播与媒体乱象，这些虚假音视频冲击着公众对“真实”本意的认知与共识。AI换脸下的虚假新闻更是消解了新闻报道内容的真实性和公众对于新闻的理解真实，同时也挑战着媒体公信力，带来前所未有的新闻伦理风险。

（一）视觉失真

智能时代下AI换脸技术对人脸信息的伪造造成视觉上的失真效果，挑战了真实性原则。媒介技术的发展增强了人不断认识世界的信息功能，拓展了人类的认知边界，可以说，技术的重要功能便是增强人们正确认识现实世界的能力。AI换脸技术的使用却是对真实人脸信息进行伪造、合成虚假图层的过程，利用人脸识别技术与深度学习的算法伪造人脸的生物特征，达到一般公众无法辨别真伪的程度，并且呈现出极具冲击性的视觉效果，具有“伪造”的技术属性。从实质上看，AI换脸技术作为一种数字工具的存在深刻影响着信息真实的建构过程，技术手段制造失真的假象反噬真实的信息本源，这正印证了麦克卢汉“媒介即讯息”的观点。

AI换脸技术将公众置于虚实难辨的信息场景之中，打破“人脸即真实”的社会共识，造成认知层面上的“失真”。隐私场景理论认为，信息流动具有情景化的特点，信息组合者将信息从原本合适的场景中抽离后将其嵌入到信息主体陌生的场景中，信息主体便无法对新场景作出准确的判断。[5] AI换脸音视频在媒体平台上的广泛传播将公众从原先真实人脸的场景中抽离，切换至真实与虚假人脸难以辨清的新场景，逐步打破了公众以人脸作为真实凭证的传统认知。而智能时代下的电子证伪机制却尚未建立，在动态的信息传递过程中公众缺乏技术手段辨别真伪，也就无法建立新的关于“真实”的认知，被动地陷入“失真”状态。AI换脸技术造成的“视觉失真”也即数字时代中“真实”的具体内涵在新技术冲击下不断消解的体现，而AI换脸技术应用于新闻行业也对新闻真实产生了新的挑战。

（二）新闻失实

学界对“新闻真实”其原则性理念基本达成一致，即新闻真实是新闻的灵魂，但对于新闻何以谓真、何以为真却莫衷一是。近年来在智能传播的大背景下出现了诸多关于“新闻真实”概念内涵的重新讨论，包括“报道真实”“假设真实”“见证真实”等说法。笔者认为，新闻真实首先应当是新闻报

道真实，是新闻报道过程中应当遵守的职业伦理准则。AI 换脸技术在媒体领域的应用中对新闻内容进行了伪造，而对 AI 换脸音视频的使用不当也破坏了新闻报道程序的正当性。

应用 AI 换脸技术制作虚假新闻的行为严重违背了新闻真实原则。真实性原则既是一种道德伦理，也是新闻行业的报道编辑原则，以及记者在报道过程中应当遵循的职业规范。在真实性原则的指导下，媒体应以真实、准确的态度报道事实，挖掘事实真相。而新闻媒体行业对 AI 换脸技术的应用却违背了真实性原则。例如 2018 年加蓬政府发布总统在电视中“向国民发表新年致辞”的换脸伪造音视频[6]，严重违背了媒体行业的真实性原则。当 AI 换脸技术被媒体用以制作虚假新闻，伪造不存在的新闻内容，真实性原则也就荡然无存。

对 AI 换脸音视频内容的不当使用破坏了新闻报道程序的正当性。“程序正义”是一种过程的正义，着眼于正义的普遍形式，考虑的是程序的正当性。[7] 实现程序正义是对新闻真实的更高要求，报道事实和报道程序正当且准确都是构成新闻真实缺一不可的要素。[8] 据 POLITCO 报纸报道，在特朗普宣布退出全球气候协定后，比利时某政党组织利用深度伪造技术制作了一段“特朗普”宣称“比利时也应该跟随美国退出全球气候协定”的 AI 换脸音视频。[9] 该新闻在报道中使用了虚假的 AI 换脸内容，并没有明显标识视频内容为合成作品，仅在视频最后作出提示，既伪造了虚假的报道内容，也在报道程序上存在不正当的操作。

（三）挑战媒体公信力

真实性原则是媒体的公信力所在。新闻真实不仅停留于新闻内容的客观真实，更强调新闻的“信任性真实”，即受众理解、信任的真实。[10] 源于社交媒体的社交性与连接性属性，有学者进一步提出了“假设真实”的概念，认为“新闻真实”在社交媒体时代更体现为媒体追寻受众信任与认可的进程，这也是媒体的公信力所在。[11] 而 AI 换脸技术对于媒体公信力的破坏，就在于其制造的虚假新闻在新媒体平台上快速扩散后引起公众对于媒体的不信任感。

自媒体为了攫取流量，大范围扩散传播充斥猎奇色彩的虚假新闻，受众因无法辨别真伪而误信，而在核查后发现自己被蒙骗，由此降低对原本报道该新闻的媒体的信任度。澳大利亚学者研究发现，在第三方机构对新闻报道中政治人物的“声称”进行核查后反而降低了受众对于进行新闻报道媒体的信任度。[12] 专业（机构）媒体核查合成新闻或制作 AI 换脸音视频警醒公众[13]，一方面使得公众将信任投资转向权威的新闻媒体，形成新闻行动者网络中不同新闻权威的分层；但另一方面自媒体同样属于新新闻生态系统的重要一环并且在数量上占据绝对优势，公众对自媒体信任度的降低反而降低了对新闻业的系统性信任。

在媒介技术的发展之下，专业（机构）媒体、自媒体、平台媒体等构成新新闻生态系统，多元的新闻行动者实质上代表了新兴而丰富的新闻采写与传播模式，但这也带来了难以统一的行为逻辑与规范理念。面对 AI 换脸技术的挑战，不同的新闻行动者都应当加强对于 AI 换脸虚假新闻的治理力度，公众也应当提升对虚假内容的辨别能力，携手共建真实、客观的社会公共舆论场与网络媒体生态环境。

五、AI 换脸的多元规制

AI 换脸技术在快速发展过程中潜藏着法律与新闻伦理风险，需要加强法律监管力度并通过媒体平台对其乱象进行规制，在市场、技术和规则之间找到平衡点。

（一）增强个人生物识别信息的保护

1. 落实人脸生物识别信息采集中的“知情—同意”原则

要严格把控人脸生物识别信息的采集，落实个人对于 AI 换脸等深度伪造技术采集生物识别信息的知情同意。根据《规定》第十四条第二款，提供显著编辑人脸、人声等生物识别信息功能的深度合成

服务提供者，应当明确提示深度合成服务使用者需要提前依法告知被编辑信息的个人主体，并获取被编辑信息的个人主体的“单独同意”而非概括同意。依据该条款，被编辑信息的个人主体仅仅能知晓其个人生物识别信息被编辑，告知不充分，个人在事先很难准确预测进一步处理人脸识别信息的“目的、方式和范围”以及使用的深度、广度，很难达到充分的“知情同意”。[14]《个人信息保护法》对收集生物识别信息等敏感信息有明确规定，在应用AI换脸技术过程中对于人脸生物信息的采集需要符合特定目的和有充分必要性，且必须告知个人信息收集的必要性，以及对于个人的影响后才能进行。由此，（深度合成服务提供者）还需要明确告知信息权利人其收集处理的内容、目的、方式及是否与第三方共享，以此实现充分知情基础上真正的同意。

2. 提高对个人生物识别信息的收集与存储标准

作为敏感个人信息，人脸识别信息的存储需要明确存储级别和措施。[15]《规定》第十四条第一款指出，深度合成服务提供者应当采取必要措施保障数据安全，确保数据在处理过程中“合法、正当”；加强数据集训练管理，提高对于涉及个人信息数据的训练数据集的保护力度，不得非法处理个人信息。依据《网络安全法》，个人生物识别信息的存储必须与个人身份信息分开存储。① 因此，在AI换脸技术的应用过程中应当针对后续人脸生物信息的存储、运输传递及使用作出明示说明，保障个人信息的安全。但《规定》中对于应用深度合成技术收集敏感个人信息后采取何种级别的保存措施，以及能否有权委托第三方保存等具体问题，尚未作出特殊规定与说明。

《民法典》第一千零三十四条中将公民个人信息分为“私密信息”和“一般信息”两类。有观点认为个人隐私及私密信息强调私密性，以及与个人尊严的强关联关系，人脸生物识别信息虽然敏感却并不私密，相反甚至具有公开性。[16] 但笔者认为，正如身高、体重、女性三围等基于可视的身体特征而提取出的以数据形式呈现的数据集一样，基于人脸生物特征提取的人脸生物识别信息同属于隐私中的私密信息，具有私密性。人脸生物识别信息集高度敏感性及私密性于一体。（深度合成服务提供者）应当分而治之地采取不同规格的保存措施存储不同层级的个人信息，将一般信息与敏感信息进行分级处理，对人脸生物识别信息等敏感个人信息采取高规格的保护措施。虽然民法典中明确规定未获得信息权利人同意不得将个人生物识别信息转移至第三方机构，但将深度合成服务提供者所承担的采集个人生物识别信息，以及存储个人生物识别信息功能相分离，引入第三方存储机构有利于提高个人生物识别信息的储存安全系数。

（二）加强媒体平台对深度合成作品的管理

1. 强化平台的主体责任

《规定》第七条、第八条中明确指出深度合成服务提供者的主体责任与应尽义务，如完善平台与创作者的服务协议，在内容上传前要求创作者自觉对内容进行标识，并采取技术或者人工方式对深度合成服务使用者的输入数据和合成结果进行审核，对于未进行标识的内容应当停止传输；以显著方式提示深度合成服务使用者承担相应的信息安全义务等。因此，作为深度合成服务的提供者，同时具有舆论属性与社会动员能力的媒体平台，必须承担相应的主体责任。

我国新媒体中的AI换脸音视频普遍存在多媒体平台传播的现象，当AI换脸视频制作完成后传播至微博、微信等其他平台时，就可能存在多方的深度合成服务提供者。此时面对多个主体承担连带责任的情况，承担责任的深度合成服务提供者是否需要从内容制作平台追溯至其他传播平台存有争议。依据《规定》第二十三条，提供深度合成服务的组织、个人属于深度合成服务提供者，为深度合成服务提供技术支持的组织、个人属于深度合成服务技术支持者。显然，传播平台应当属于“深度合成服务

① 国家市场监督管理总局、国家标准化管理委员会正式发布国家标准《信息安全技术 个人信息安全规范》第5.4条规定。

提供者”，AI 换脸技术的使用者并非孤立地散播 AI 换脸深度合成音视频，其往往借助媒体平台发布，以达到大范围传播扩散的效果。因此，笔者认为，无论作为内容制作的平台，抑或由平台用户转发扩散的其他传播平台，都应该承担相应的责任。《规定》第六条：“不得利用深度合成服务制作、复制、发布、传播虚假新闻信息。”法律明确禁止媒体平台及用户利用 AI 换脸技术合成虚假新闻。媒体平台及用户不得利用 AI 换脸技术杜撰虚假的新闻内容，无论是对个人或名人的换脸或对其面部微表情进行转换等，都是违法的。随意使用国家领导人的肖像进行深度合成，在我国不但构成侵权，还存在其他风险。另外，如果转载的深度合成的新闻作品中涉及他人肖像，如虚拟主播主持的新闻作品、VR 新闻等，应当“依法转载互联网新闻信息稿源单位发布的新闻信息”。

2. 深度合成作品必须作出明显标识

依据《规定》第九条、第十七条①，作为深度合成服务提供者，平台必须要求信息提供者进行身份认证并对其上传的相关的深度合成内容作出明显标示标识，促进塑造客观、真实的媒体生态环境。平台可以应用区块链技术为用户提供在深度合成作品上作出明显标识的技术手段，如运用数字水印、分布式共识和经济激励等手段，实现去中心化信用的点对点交易与协作。[17] 平台还可以在内容中应用溯源防伪技术增强网络平台中对用户隐私的保护，如一种可以应对基于 GAN 的编辑与篡改的端到端的深度学习方法，解决原创者作品被肆意伪造修改而难以追溯源头的难题。[18] 阿里云开发者社区建立了时间戳在线转换工具，能够为用户提供一份电子证据以证明用户的某些数据的产生时间，普通公众可以通过保存时间戳作为“数字水印”保护个人创作作品，积极配合平台管理深度合成内容。

除建立明显标识的保护之外，平台也应当防范出现深度合成内容未进行明显标识的情形。《规定》第十八条，任何组织、个人不得采用技术手段删除、篡改、隐匿相应的深度合成标识。当平台检测到信息提供者未对规定的深度合成信息内容进行显著标识时，应当立即中止内容的发布，并要求上传者按照规定标识显著记号才能进行公开与传播。当平台检测到未按规定进行显著标识的深度合成信息内容在其平台上广泛传播时，应当立即封锁该信息的阅读查看与转发权限，并对内容传播者进行通告处理。

3. 增强对深度合成内容的监测

平台需要不断引进新型检测技术，扩展新的组件技术提高检测精确性，以解决新问题。当下针对 AI 换脸的伪造音视频已有专门的探测技术进行检验，包括多种探测方式：其一，从人脸固有特征发现生理特征上的区别；其二，从伪造方法出发，观察生成对抗网络的固有痕迹或通过修改卷积神经网络结构来检测伪造视频；其三，抓住视频内容的时序性进行伪造检测。目前绝大多数方法的准确率指标都已达到 95%，但是这些指标都是在公开数据集上得到的，实际互联网环境中的准确率只有 70%左右。[16] 而类似于能够跟踪用户所有对话、行为轨迹的生命日志记录又存在明显的侵犯用户个人隐私的弊病，难以真正推行。因此，以上技术都并非尽善尽美，平台亟须在实际应用中不断革新检测技术，加强对深度合成内容的检测管理。

平台还应当加强对深度合成内容的审核。建立健全用于识别违法和不良深度合成信息内容的数据库，提高对违法和不良深度合成内容的识别度与检测准确率，对于其中违法以及不良的内容进行删除。当发现因审核失误而导致违法和不良的深度合成内容扩散传播时，平台应立即封锁该信息的阅读查看与转发权限，以中止虚假信息的传播，对相关的信息提供者予以处置，并及时采取相应的辟谣措施，

① 《规定》第十七条，“深度合成服务提供者提供以下深度合成服务，可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置、区域进行显著标识”中列举五种情况，其中第三款即 AI 换脸“人脸生成、人脸替换、人脸操控、姿态操控等人物图像、视频生成或者显著改变个人身份特征的编辑服务”。

向用户公示虚假内容及比对情况，同时将相关信息报相关部门备案，承担起相应的管理职责。

（三）提高公众的媒介素养

1. 提高信息甄别能力

英国学者汤普森和利维斯认为，媒介素养就像“疫苗”，大众通过接种“疫苗”便能够对当时通俗流行文化“免疫”，从而保持独立思考、批判的精神。[19] 当下以技术为主导的媒介逻辑重构着社会的各个领域，AR、VR、直播等技术手段建构起融合多重感官的沉浸式媒介体验，大众亟待提升对庞杂而富媒化的信息内容进行甄别和批判性思考的处理能力。在面对AI换脸等合成技术制造的虚假新闻时，一方面，公众应在视觉冲击前保持冷静理智的头脑，注重媒介内容的客观性、真实性、准确性，弱化主观性、情感性等非理性因素的影响，时刻保持对媒介信息的批判意识；另一方面，公众应当理性甄别媒介信息，对带有深度合成标识的内容保持警惕，对于权威信源投以更多的信任而非将关注放在小道消息与营销号等自媒体消息上，不盲从、不盲信媒介信息内容。

2. 增强媒介使用能力

在万物皆媒的数字时代，网络媒体生态在技术赋权下，不仅是投射真实世界的拟态环境，而且在各类人机交互的社交网络中虚实相融、信息杂糅，对当代公民素养提出了更高的要求。“技术是一种解蔽方式”，海德格尔在《技术的追问》中提出应以技术手段破解技术困境。如何鉴别AI换脸技术中的虚假新闻，也需要公众和平台积极、有效地利用一些检测网站以及事实核查平台对新闻内容进行核查，主动使用相关媒介技术了解技术黑箱的运行法则，作出理性分析与判断。在媒介使用过程中，还应当对媒介技术所产生的技术风险有所防范，尤其注意个人生物识别信息安全，涉及人脸识别、人声识别、指纹、瞳孔等生物识别信息时需要充分了解信息的采集、使用与存储流程以及信息采集目的的充分必要性以及使用情况，谨慎使用生物识别信息。公众应积极使用媒介技术发展“延伸”自身的信息功能，在新媒体的使用过程中开拓视野、培养多元化思维以及建构信息知识网络，在虚拟的网络世界中提升自身的信息获取能力、信息使用效率，主动参与到社会公共事务的讨论与民主决策中去，成为理性的网络公民。

3. 树立法制观念

媒介素养极具情境性，伴随着信息传播技术的快速变化，其内涵与维度也发生着不同转变。在AI技术与社会化媒体深度融合发展下，智能技术改造升级传统数字传播语境的同时也带来一系列法律风险，诸如使用AI换脸侵害肖像权、创作剪辑类短视频侵害著作权，以及个性化算法推荐过度收集个人信息侵害隐私权等。这要求公民具备一定的与媒体相关的法律知识与法制观念。“网络空间并非法外之地”，公众在使用AI换脸技术制作娱乐类音视频的过程中要注意其带来的法律风险。公众应当学习相关的网络安全法律法规，增强法制观念并将其内化为自律行为，共同构建法治化的媒体生态环境，清朗网络空间。

六、结 论

“走太快了，灵魂跟不上。”这是一个游牧部落的古训，也是对人工智能技术突飞猛进但隐患频仍现状的映射。我国AI换脸技术的开发和利用正处于高速发展阶段，但对其产生的技术风险以及必要的规制对策还缺乏足够的认识和深入的研讨。AI换脸技术是一把利弊兼具的双刃剑，在深度合成音视频娱乐大众的同时，其技术滥用而缺少监管也潜藏着生物信息安全隐患与传播侵权的局部无序乱象。而AI换脸技术应用于新闻制作，尤其涉足政治领域的严肃新闻时，加剧视觉失真与新闻失实的困境，形成复杂的网络舆论生态混沌。

加快规范技术开发的立法，让技术插上伦理的翅膀，乃是当务之急。同时，要积极预防新媒体AI

换脸技术应用对社会信任机制的冲击，加强事实核查机制的建设，建立健全谣言核查平台与辟谣机制。对于 AI 换脸等深度合成技术的应用，应该在个人信息安全、企业基于用户信息的技术创新发展、数字社会治理上找到平衡。

参考文献：

[1] Brinckerhoff，R. （2018）. Social network of social nightmare：How California Courts can prevent Facebook’s frightening foray into facial-recognition technology from haunting consumer privacy rights forever. *Federal Communications Law Journal*，70（1）：105-106.

[2] 张煜之，王锐芳，朱亮，赵坤园，刘梦琪. 深度伪造生成和检测技术综述［J］. 信息安全研究，2022，8（3）：258-269.

[3] 龙坤，马钺，朱启超. 深度伪造对国家安全的挑战及应对［J］. 信息安全与通信保密，2019（10）：21-34.

[4] 孙道锐，张丽洁. 利用信息技术手段伪造侵权的法教义学分析［J］. 法律适用，2021（5）：166-176.

[5]［美］海伦·尼森鲍姆：《信息时代的公共场所隐私权》［A］. 张民安.《公共场所隐私权研究》［C］. 广州：中山大学出版社，2016：82-83.

[6] Faife，C. In Africa，fear of state violence informs deep fake threat. Retrieved December 09，2019，fromhttps：//blog. witness. org/2019/12/africa-fear-state-violence-informs-deepfake-threat/

[7] 姚大志. 论程序正义［J］. 天津社会科学，2000（4）：39-42.

[8] 窦锋昌. 新闻真实有赖于程序真实［J］. 青年记者，2022（14）：127.

[9] Burchard，H. Belgian socialist party circulates ‘deepfake’ Donald Trump video. Retrieved December 11，2022，from https：//www. politico. eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

[10] 李唯嘉. 如何实现“信任性真实”：社交媒体时代的新闻生产实践——基于对 25 位媒体从业者的访谈［J］. 国际新闻界，2020，42（04）：98-116.

[11] 操瑞青. 作为假设的“新闻真实”：新闻报道的“知识合法性”建构［J］. 国际新闻界，2017，39（5）：6-28.

[12] Carson，A.，Gibbons A.，Martin A. &Phillips，J. B. （2022）. Does third-party fact-checking increase trust in newsstories? An Australiancases tudy using the“sports rorts”affair. *Digital Journalism*，10（5）：801-822.

[13] 澎湃新闻.《【智库观点】新技术安全：深度伪造技术的应用场景与风险》［EB/OL］.［2021-01-27］. https：//www. thepaper. cn/newsDetail_ forward_ 10950428

[14] 程啸. 论个人信息处理中的个人同意［J］. 环球法律评论，2021，43（6）：40-55.

[15] 倪楠，王敏. 人脸识别技术中个人信息保护的法律规制［J］. 人文杂志，2022，（02）：121-131.

[16] 林凌. 人脸识别信息“人格权-用益权”保护研究［J］. 中国出版，2021，（23）：41-46.

[17] 孙毅，王志浩，邓佳，李犇，杨彬，唐胜. 人脸深度伪造检测综述［J］. 信息安全研究，2022，8（3）：241-257.

[18] 王丽娜，聂建思，汪润，翟黎明. 面向深度伪造的溯源取证方法［J］. 清华大学学报（自然科学版），2022，62（5）：1-6.

[19] Chen，R-T.，Wu，J. & Wang，Y. M. （2011）. Unpacking new media literacy. *Systemics*，*Cybernetics and Informatics*，9（2）：84-88.

［责任编辑：谢薇娜］