人工智能深度伪造技术的法律风险防控

龙 俊 王天禹

[摘 要] 作为生成式人工智能的深度伪造技术给社会带来了新的风险与挑战。从类型学上看，深度伪造技术同时引发了基于个体权利谱系的微观法律风险、基于市场经济秩序的中观法律风险和基于国家公共安全的宏观法律风险，并依此形成了多样化和梯次性的法律风险格局。为有效化解风险，各国存在“自下而上型”分散治理模式、“自上而下型”垂直治理模式以及“多方参与型”合作治理模式。我国可通过“软法+硬法+算法”相结合的综合治理模式，实现对前述法律风险的有效防控：在软法层面，完善科技伦理制度，明确科技伦理的实体规范和程序规范；在硬法层面，通过“风险预防”与“风险控制”，强化风险防控的刚性约束；在算法层面，实现从样本数据采集到深度伪造内容发布的全过程技术监督，发挥技术的规范功能。

[关键词] 深度伪造；破坏性创新；法律风险；科技伦理；算法
[中图分类号] D63 [文献标识码] A [文章编号] 1674-7453（2024）03-0069-11

近年来，深度学习（Deep Learning）技术的迅猛发展推动了人工智能领域的重大创新，作为生成式人工智能的深度伪造[①]技术则是通过利用多种神经网络模型来生成、合成文本、图像、音频、视频等内容。随着深度伪造技术的广泛应用，其背后的法律风险也日益凸显。尽管世界各国（地）已经采取了一系列措施来应对这些风险，但现有的治理模式仍然存在一定缺陷。目前学界关于深度伪造技术的治理研究大多聚焦于宏大叙事，缺乏系统、分层、有针对性的研究成果。在此背景下，本文试图厘

①一些学者和业内人士认为“深度合成”的表述更能表明技术中立性，但本文侧重分析“深度伪造”技术引发的法律风险，因而延用“深度伪造”这一表述。参见曹建峰：《AI生成内容发展报告2020—“深度合成”（deep synthesis）商业化元年》，https://tisi.org/14419，20-12-2023.

[基金项目] 2021年度福建省高校以马克思主义为指导的哲学社会科学学科基础理论研究项目“优化营商环境与深度伪造技术法律风险防控研究”（JSZM2021023）。

[作者简介] 龙俊，福建师范大学法学院副教授、硕士生导师；王天禹，福建师范大学地方治理与地方法治研究中心助理研究员。

清深度伪造技术的基本原理与效应，系统分析深度伪造技术的法律风险类型，在考察各国现有治理模式的基础上，从伦理、法律和技术三个维度提出具体的对策建议，以有效防控深度伪造技术带来的法律风险。

一、深度伪造技术的基本原理与应用效应

在充满“不确定性”的科技时代，[1]任何技术风险的防范首先依赖于对技术本身运作方式和场景的剖析和解构。因此，系统阐明深度伪造技术的基本原理及其应用效应，是有效防控深度伪造技术法律风险的逻辑起点。

（一）技术原理：从延续性创新到破坏性创新的历史流变

从技术演进与创新理论的双重视角来看，深度伪造技术本质是破坏性创新的结果。20世纪40年代，经济学家约瑟夫·熊彼特（Joseph A. Schumpeter）曾提出创造性破坏（Creative Destruction）理论，指出创新是经济增长的内生动力，包括新技术、新产品、新市场等要素，[2]其中技术革命是创新的关键驱动。在此基础上，克莱顿·克里斯坦森（Clayton M. Christensen）进一步阐释了这一创新理论，将技术创新分为延续性创新（Continuous Innovation）与破坏性创新（Disruptive Innovation），[3]前者是指在现有技术的基础上进行改进或提升，后者是指打破现有的市场和技术，开发出全新的产品，当延续性创新的方式难以有效刺激、满足和服务实践需求时，破坏性创新便应运而生。[4][5]在人工智能领域，“深度伪造”的出现就是深度学习技术从延续性创新到破坏性创新的典型例证。与传统神经网络模型不同，深度伪造技术主要采用生成对抗网络（Generative Adversarial Network，GAN）架构，[6]（Generator）生成器负责生成逼真的样本数据，而判别器（Discriminator）则负责判断生成的样本数据与真实样本数据的区别，在二者相互对抗的过程中，模型性能和学习能力不断得到提升。

具体而言，GAN的破坏性创新在于其处理数据的方式和目标与以往人工智能模型相比具有颠覆性差异。一方面，GAN作为生成模型无需面临传统模型中指数级上升的计算量。由于GAN生成数据的复杂度与数据维度①呈线性相关，故对于较大维度样本的生成仅需增加神经网络的输出维度，同时，生成模型对数据分布不作显性限制，②省去了人工设计模型分布这一环节。另一方面，GAN是兼具监督式学习和非监督式学习双重特征的神经网络架构。在模型训练过程中，生成器通过学习训练数据生成新的样本，并与判别器提供的反馈相结合来进行训练，[7]这个对抗过程可以被视为一种特殊的监督式学习，其中判别器的输出被视为监督信号。基于这两方面优势，GAN成为处理大量无标签数据的高效工具，能够在各种任务中生成逼真的样本。在这个意义上，深度伪造不再是对以往深度学习技术的简单改进，而是实现了突破性进展。

（二）技术效应：积极效应与消极效应并存

由于技术应用的便捷性与伪造内容的仿真性不断提高，深度伪造技术被广泛应用于诸多场景，并同时产生了积极与消极效应。

---

①数据的维度是指数据中的特征数量或者变量数量。例如，在图像生成模型中，图像的维度通常是指图像的宽度、高度和通道数。在文本生成模型中，数据的维度可能指的是文本的长度或者一个句子中的词数。

②数据分布的显性限制是指对数据的一种约束条件，通常用于描述数据的某些特性或限制条件。在机器学习和数据科学中，显性限制可以用于指导模型的设计和训练，以确保模型能够适应并符合这些限制条件。

1. 产生积极效应的场景

深度伪造技术在诸多领域都产生了显著的积极效应。在科研领域，深度伪造技术可以帮助科研人员在收集数据之外，还能进行高质量的数据训练，或者通过深度伪造模拟完成一些具有高风险的试验，如利用自动驾驶仿真系统（AADS）完成测试。在教育领域，深度伪造为教育提供了崭新的技术工具，使教育内容能够获得全方位展示，如通过深度伪造使历史人物“重现”，让损毁或消失的建筑（文物）得以“复原”。在文化娱乐方面，深度伪造复刻、重组现实的能力可以丰富影视、游戏在声音和视觉上呈现的效果，提升了内容的趣味性。在医疗卫生方面，深度伪造技术不仅能生成与医学影像无异的数字影像供医生参考，有时甚至可以直接介入治疗和康复的过程。此外，在其他商业领域，如社交网络、电子商务等方面，深度伪造可以为个体（消费者）创造虚拟化身（Avatar）参与各类社交（消费）活动，或创造虚拟主播、虚拟偶像以及在各个平台上生成与之服务相匹配的虚拟形象与用户进行交互等。

2. 产生消极效应的场景

虽然深度伪造技术释放了巨大的应用潜力，但技术滥用产生的消极影响同样不容忽视。在私人领域内，利用深度伪造技术仿冒他人身份实施违法犯罪行为的现象越来越常见。在商业活动中，利用深度伪造技术对影视、音乐等作品进行篡改、贬损也是较为普遍的情形。随着用户生成内容（User-Generated Content）逐渐成为互联网内容生产的重要方式，通过深度伪造对原创作品进行二次创作的用户也日益增多。然而，创作行为本身对著作权和市场秩序可能产生的侵害同样不容忽视。这些利用深度伪造技术传播各类虚假信息的行为不仅侵犯了相关主体的合法权益，而且也对社会秩序的稳定不利。

二、深度伪造技术应用的法律风险类型分析

为消弭深度伪造技术带来的消极效应，有必要将技术应用中涉及的法律风险作类型化处理，以便决策主体建立更为清晰的认知模型，降低法律风险防控的决策成本（如图1所示）。

（一）类型Ⅰ：基于个体权利谱系的微观法律风险

1. 第一重法律风险：侵犯公民人身权利

深度伪造技术最早流行的应用场景就是将原始素材伪造生成具有欺骗性的视频、音频、图像等内容，而这些行为会对公民各项人身权利造成威胁。例如，在人脸信息获取方式便捷多样的移动互联网时代，如果行为人在未经授权的情况下通过深度伪造技术获取并使用公民肖像，则可能侵害公民的肖像权。又如，在深度伪造技术的应用过程中，若施害者在获取原始素材时不当干扰公民享有的私人生活安宁与私人空间，侵扰、利用或公开个人隐私，则构成对公民隐私权的侵犯[①]。

2. 第二重法律风险：侵犯公民财产权利

深度伪造还可能成为侵犯公民财产权利的新型工具，帮助不法行为人实施《中华人民共和国刑法》（以下简称《刑法》）中的各项财产类犯罪。一方面，行为人可以通过盗取他人信息直接实施犯罪活动，如通过伪造人脸信息破解人脸识别密码，盗用虚假身份创建银行账户实施诈骗等。另一方面，行为人还可能通过伪造他人不良信息间接实施犯罪活动，例如通过伪造各类不利于受害者的虚假信息对其进行敲诈勒索。由此可见，深

① 参见《中华人民共和国民法典》（以下简称《民法典》）第1019条、第1024条、第1032条。

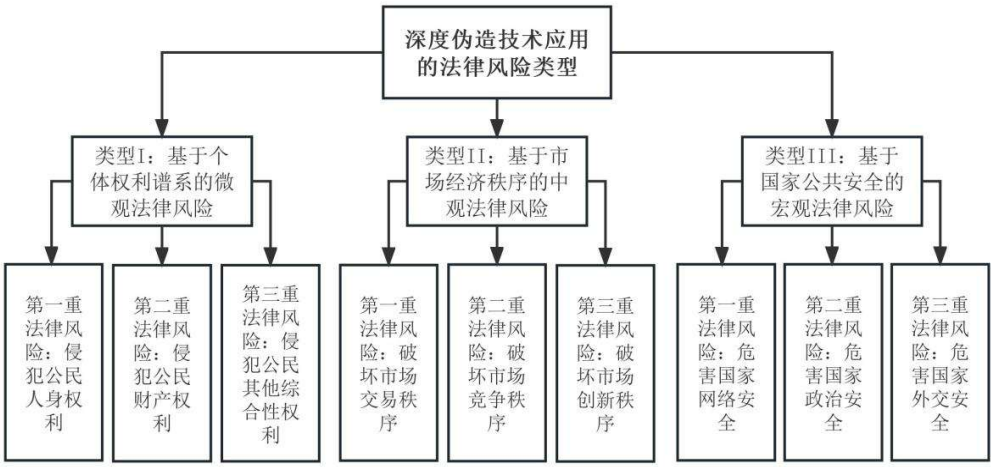

图1 深度伪造技术应用的法律风险类型图

度伪造技术的应用一定程度上加剧了公民财产权利遭受侵犯的风险，为各类财产型犯罪提供了空间。

3. 第三重法律风险：侵犯公民其他综合性权利

深度伪造技术在应用过程中不可避免会影响个体的信息权益或数据权利等新兴综合性权利，从而给《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）《中华人民共和国数据安全法》（以下简称《数据安全法》）的实施带来挑战。一方面，目前法律规定的“知情同意”规则在具体落实上尚不能满足用户的实际要求，用户在使用与深度伪造相关的应用软件时，可能面临个人信息被不当收集、使用、加工、传输的法律风险。另一方面，在滥用深度伪造技术的场景中，只要伪造内容一经发布和传播，内容中所承载的个人信息和数据即可在网络空间内被随意使用，从而导致个人信息泄漏或数据权益遭受损害的法律风险。

（二）类型II：基于市场经济秩序的中观法律风险

1. 第一重法律风险：破坏市场交易秩序

在市场交易活动中，主体适格、意思表示真实以及内容合法是经营主体之间实现有效交易的基本法律要件，但深度伪造技术的滥用则可能对前述正常的市场交易活动造成阻碍。例如，行为人可能通过伪造身份展开市场交易，掩盖主体资格瑕疵或影响交易方的真实意思表示，进而使得交易的合法性问题产生难以预估的法律风险。又如，在金融交易市场上，有的经营主体可能利用深度伪造技术散布伪造的金融交易信息等。这些行为都可能严重危害市场交易安全、破坏市场交易秩序。

2. 第二重法律风险：破坏市场竞争秩序

深度伪造技术在商业领域的应用为经营者实施虚假宣传、商业诋毁等不正当竞争行为提供了技术手段，进而对《中华人民共和国反不正当竞争法》所保护的市场竞争秩序造成损害。例如，经营者在商业宣传中利用深度伪造技术对商品作出与实际不符的虚假描述，引发消费者误认，或者通过深度伪造技术编造、传播虚假信息，损害其他经营者的商业信誉等，对市场竞争秩序造成更为隐蔽的破坏。一方面，深度伪造技术具有通用性特征，经营者在任意时刻和环境都可利用该

技术实施商业诋毁等不正当竞争行为，从而加大了确定施害主体的难度。另一方面，虽然视频已成为当下最能提供可信内容的载体之一，但深度伪造技术却颠覆了视频内容的可信度，由此产生的不正当竞争行为更使得消费者的认知错误在短期内难以纠正，[8] 进而大大提高了市场的辟谣成本。

3. 第三重法律风险：破坏市场创新秩序

由于通过深度伪造技术进行作品创作往往是基于拼接、替换与合成等方法完成的，并且创作内容与原生内容高度重合，故该技术的使用在一定程度上可能打击原创的积极性，进而破坏市场创新秩序。从实定法上看，深度伪造技术为现行《中华人民共和国著作权法》的实施带来了新的挑战。其一，在“二次创作”的著作权归属问题未有定论的背景下，将具有强大二次创作能力的深度伪造技术广泛应用于内容生产，则无疑会对著作权规则造成冲击。其二，对于利用深度伪造技术进行再创作是否可以提出合理使用的要求问题，不仅涉及衍生作品的利益分配，而且直接影响着市场创新的激励机制。如果禁止提出合理使用的要求，可能会阻碍深度伪造技术在创作领域的应用前景。可见，深度伪造技术的广泛应用也可能在著作权领域引发破坏市场创新秩序的法律风险。

（三）类型 III：基于国家公共安全的宏观法律风险

1. 第一重法律风险：危害国家网络安全

网络空间是传播深度伪造内容的主要渠道，深度伪造技术的滥用可能会危害国家网络安全。可能的情形包括：通过深度伪造将社会事件过度夸张渲染，捏造、歪曲新闻事件中的事实以博取公众眼球、引导社会舆论、赚取用户流量，发布伪造的灾情、险情、疫情等信息，妨碍救助工作的展开，等等。

2. 第二重法律风险：危害国家政治安全

利用深度伪造的欺骗性实施违法犯罪行为，可引发对《刑法》具体法益的侵害，从而产生对应的法律风险。例如，以深度伪造技术实施或煽动实施分裂国家、破坏国家统一、颠覆国家政权、推翻社会主义制度的行为，可能构成分裂国家安全罪、煽动分裂国家罪等罪名；以深度伪造技术窃取、篡改、传播国家机密文件或数据信息，可能构成泄露国家秘密罪等①。

3. 第三重法律风险：危害国家外交安全

信息时代各国的联系日益紧密，国家在外交方面亦需格外注意防范深度伪造技术潜在的法律风险，避免违反国际法的基本原则。例如，在国际交往中滥用深度伪造技术可能会破坏国际交流的信息环境和国家之间的相互信任，[9] 导致国家之间产生误解，从而影响相关国际义务的履行。

三、深度伪造技术现有治理模式的利弊考察

为应对深度伪造技术带来的法律风险，世界各国（地）纷纷出台相关政策举措。但由于各地历史文化、法律传统、技术水平等因素的不同，各类治理路径亦呈现出不同特点，且各有利弊。

（一）域外的典型治理模式

域外对深度伪造技术的治理主要包括美国模式和欧盟模式两种。其中，美国高度重视市场调节的功能，认为政府作为“守夜人”通常只有在新兴技术对社会造成明显负担后才被动采取回应措施，[10] 以避免过严的防控对市场造成不当侵害，故形成了“自下而上型”分散治理模式。该模式遵循“市场自治为主、政府管制为辅”的理念，通过企业自审、技术研发、法律助推等“自下而上”的系列措施实现风险管控。然而，由于该模式过

①参见《刑法》第103条、第105条、第398条。

表1 我国“多方参与型”合作治理模式

治理路径	政策举措	具体内容
中央层面加强顶层设计	· 2020 年 12 月中共中央印发《法治社会建设实施纲要（2020-2025）》	· 制定和完善对网络直播、自媒体、知识社区问答等新媒体业态和算法推荐、深度伪造等新技术应用的
	· 2022 年 3 月中共中央办公厅 国务院办公厅印发《关于加强科技伦理治理的意见》	· 加快构建中国特色科技伦理体系，健全多方参与、协同共治的科技伦理治理体制机制，强化底线思维和风险意识，建立完善符合我国国情、与国际接轨的科
国家互联网信息办公室、公安等部门主动出台规范	· 2019 年 11 月国家互联网信息办公室、文化和旅游部、国家广播电视总局发布《网络音视频信息服务管理规定》	· 网络音视频信息服务提供者、使用者对深度合成的信息应当进行标识，不得利用深度合成制作、发布、传播虚假新闻
	· 2021 年 4 月国家互联网信息办公室、公安部等七部门发布《网络直播营销管理办法（试行）》	· 直播营销平台应当对利用人工智能、数字视觉、虚拟现实、语音合成等技术展示的虚拟形象从事网络直播营销的，应当按照有关规定进行安全评估，并以
	· 2021 年 11 月国家互联网信息办公室、工业和信息化部等四部门发布《互联网信息服务算法推荐管理规定》	· 算法推荐服务提供者提供互联网新闻信息服务的，不得生成合成虚假信息
	· 2022 年 11 月国家互联网信息办公室、工业和信息化部、公安部发布《互联网信息服务深度合成管理规定》	· 明确深度合成服务的一般性规定，数据和技术管理规范，监督检查与法律责任以及相关概念的含义
	· 2023 年 7 月国家网信办、国家发展改革委等七部门发布《生成式人工智能服务管理暂行办法》	· 明确提供生成式人工智能产品或服务应遵循的基本要求，以及需要承担的责任和应当履行的义务
各类机构积极参与治理	· 清华人工智能研究院、北京瑞莱智慧科技公司、国家工信安全发展研究中心等联合发布《深度合成十大趋势报告（2022）》	· 从技术研究、领域应用、发展趋势等多个方面，全面深入地介绍和研判深度合成技术及应用带来的机遇与挑战，并就其发展与治理给出切实可行的建议
	· 百度、360、华为、阿里等企业与上海人工智能实验室联合担任国家人工智能标准化总体组大模型专题组组长	· 积极推动深度合成等生成式人工智能领域大模型国家标准体系建设，牵头起草国际、国家或行业标准
	· 同济大学、国家新一代人工智能专委会等发布《人工智能大模型伦理规范操作指引》、《新一代人工智能伦理规范》等	· 积极探索制定各类行业规范，以构建生成式人工智能的应用管理规范、研发规范、供应规范和使用规范，将伦理道德贯穿于产业科技创新活动全过程

于依赖企业自治，引发了诸多问题。例如，虽然美国诸多新兴科技企业通过建立 AI 伦理委员会对 AI 产品进行道德伦理审查，并形成了伦理规范的约束机制，但是由于缺乏统一的法律规范，且各州法律之间极易发生冲突，[11] 所以美国至今尚未形成有效的法律风险防控协调机制。此外，美国虽然拥有强大的高科技企业群，可以对 AI 立法施加影响，但各企业之间难以形成立法共识。

相比之下，欧盟虽然没有形成针对规制深度伪造技术的专门立法，但选择了以保护基本权利、构建科技伦理为主体，同时要求社会各界采取针对性措施的“自上而下型”垂直治理模式。[12] 该模式以欧盟委员会为主导，对网络服务提供者、技术研发机构和公众分别提出了不同要求，以应对新兴技术带来的风险。欧盟模式深受欧洲本土历史传统的影响，例如在法治理念上着重人权保障，尤其对个人基本权利的保护；[13] 在治理逻辑上更强调能动治理，以尽早化解技术带来的各类风险。该模式同样存在明显缺陷。例如，欧盟成员国对 AI 技术的治理态度存在分歧，民众意见难以通过欧洲议会获得充分表达。又如，欧盟的“强监管模式”对技术本身的创新和发展产生了一定的“压制”作用，且未能充分调动社会主体主动参与技术治理的积极性。

（二）我国的“多方参与型”合作治理模式

我国历来重视网络安全保障，对深度伪造技术的法律风险防范也一直采取积极应对的态度，不仅出台了各类规范性文件，而且在借鉴世界各地实践经验的基础上，探索出了独具特色的“多方参与型”合作治理模式（参见表1），中央层面加强顶层设计、各级机关主动出台规范、各类机构积极参与治理”的多元共治局面。从效果上看，这一模式兼顾了民主与效率，调动了各方主体的积极性，但是在具体实施中同样产生了一些问题。例如，公权力机关对包容审慎监管、常态化监管等理念尚未形成统一认知；各部门出台的法律规范层级较低，且多为原则性、倡导性的“意见”“暂行规定”“管理办法”等，缺乏足够的刚性约束；相关文件“政出多门”，体系性和协调性有待提高。

综上所述，三种治理模式在多个维度上均存在差异：自下而上的分散治理模式更侧重对自由竞争和创新的维护，以此为导向产生了由市场调节为主的治理逻辑，其缺点则在于政府缺乏主动性，立法分散且滞后；自上而下的垂直治理模式更侧重于对个体权利的保障，强调伦理价值多于创新价值，其缺点则在于各国施政态度不一、各主体参与治理的积极性不足；多方参与的合作治理模式则既注重对安全的保障又强调对技术的利用，同时倡导以包容审慎的监管理念和底线思维来实现各方利益的平衡，但在具体措施上却仍然存在规范冗杂而体系不足等缺憾，存在可进一步优化的空间。

四、深度伪造技术法律风险防控的应对之策

深度伪造技术引发的法律风险与传统作用于物理场域的法律风险相比发生了深刻变化，故与之相应的风险防控策略也应因时而变。结合前文分析，我国可尝试从软法、硬法和算法三个维度提炼风险的防控之策。

（一）软法之治：以伦理约束技术

与法律规范的强制性相比，作为软法的科技伦理因具有引导性与灵活性等特征而在规范深度伪造技术方面独具优势。伦理约束不仅有利于从源头上规避风险，而且可以通过理性商谈达成有效共识，从而增强技术应用的可接受度。因此，有必要将伦理约束嵌入到科技活动研发、设计和应用的全过程之中，通过法律规范促进伦理的反思与完善，形成层次分明的伦理规范体系。

一是在深度伪造的伦理实体规范层面，可分为宏观与微观两个角度讨论。在宏观上，遵循《新一代人工智能伦理规范》[①]，将伦理道德融入人工智能的全生命周期，为从事人工智能相关活动的自然人、法人和其他机构提供伦理指引。在微观上，一方面，根据深度伪造技术应用的具体场景进行决策，在技术研发的上游宜采用伦理约束的方式，开展软法规制，设置原则性、程序性的要求，而在技术应用的下游宜采取回应性规制，设置具体的法律责任。另一方面，依照《科技伦理审查办法（试行）》的基本精神，建立相关部门科技伦理委员会的审查标准，明确伦理审查的监管职责；尝试在科技伦理委员会下设立不同类别的专门委员会，将深度伪造的伦理审查任务纳入AI技术专门委员会监管；制定“深度伪造科技伦理指引”，指导从业机构开展科技伦理治理工作；等等。

二是在深度伪造的伦理程序规范层面，构建

①国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》，充分考虑当前社会各界有关隐私、偏见、歧视、公平等伦理关切，提出了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等6项基本伦理要求。同时，提出人工智能管理、研发、供应、使用等特定活动的18项具体伦理要求。

人工智能伦理商谈程序以达成稳定共识。建立伦理共识商谈程序能够整合技术风险涉及的相关知识主体、利益主体和价值主体，为多元主体均有机会参与到科技伦理商谈的过程提供制度平台。[14] 可以构建科学咨询程序（通过遴选具备深度合成专业知识的资深专家为制定科技伦理规范提供科学性基础）、同行评审程序（通过外部同行专家对伦理规范制定的科学依据进行评价）和公众参与程序（通过正式的听证会和非正式的公共论坛等形式为公众参与伦理商谈提供有效途径），并在技术研发阶段、应用阶段和产品上市阶段，通过相关知识主体、利益主体和价值主体参与伦理共识商谈程序，最终确立技术所须遵循的伦理规范。

（二）硬法之治：以法律规范技术

法律规范能以刚性约束的方式对各类主体的权利义务边界予以明确。从系统论思维来看，治理深度伪造技术包含“风险预防”与“风险控制”两个层面，二者的相互结合与共同优化可以为深度伪造技术法律风险的防控提供新的思路。

1. 风险预防阶段的法律优化路径

在风险预防阶段，赋予深度合成服务提供者（以下简称“服务提供者”）有限的监管权限，在监管部门的督促和引导下进行自我规制，有助于从源头上减少深度伪造技术引发的法律风险。

第一，要求服务提供者履行“守门人”（Gatekeeper）职责。目前服务提供者审查义务的范围尚不明确，扭曲了自我规制的激励机制，服务提供者为降低合规成本会过度侵蚀用户的权利。[15] 为此，监管部门应当出台明确的深度伪造内容负面清单，为服务提供者履行“守门人”职责提供指引。同时，服务提供者应定期向监管部门反馈深度伪造内容的审查情况，以便动态调整负面清单，形成良好的合作治理机制。

第二，引入技术避风港（Technology Safe Harbor）规则。技术避风港分为两种类型，一是技术白名单，即监管机构审核认定的技术方案，如果服务提供者部署白名单内的技术，就可以对技术使用者的违法内容免责；二是技术自评估，即服务提供者自行评估以证明其所使用的技术方案符合合规标准，并向监管机构提交自评估报告，一旦取得监管机构的认证，则可以就相应的违法内容免责。该规则可以为服务提供者提供免责预期，激励服务提供者在合规方面作出积极的投入。

第三，细化知情同意原则。现行立法规定了知情同意规则①，但在具体实施过程中流于形式、缺乏详细指引。一方面，在用户不同意相关个人信息条款时，将不能使用深度合成服务，实际上服务提供者将“同意”与“使用”捆绑，具有一定的强制性；另一方面，个人信息在被“合法”获取后仍会面临诸多处理流程，但服务提供者仅凭用户的一次性同意即可获得用户的个人信息权益。为此，可以考虑建立动态的知情同意机制，即在服务的各个阶段均须取得用户同意，并提供个人信息实时状态的可查询服务。在必要时，用户可以撤回“同意”，并要求服务提供者定期发布用户个人信息使用报告，从而加强个人信息处理的全过程监管。

第四，明确信息披露义务。具体包括以下三个方面。一是确立算法原理公开规则，要求服务提供者以明确和易于理解的方式向用户主动说明深度伪造算法模型的基本原理，并在用户有需要时提供可解释的合成结果。二是建立“披露核查—反馈”机制，服务提供者要利用检测技术核查披露情况，就检查情况与未履行披露义务的用户进行沟通、反馈并督促完成或辅助完成标识义务。三是提供内容提示/警告服务，在用户浏览深度伪造内容或对深度伪造内容进行“点赞”“评论”“转发”等操作时，应自动弹出相应的提示/

①参见《民法典》第1035条、《网络安全法》第41—42条、《个人信息保护法》第13—15条。

警告窗口告知用户内容为深度合成，以符合用户预期。

第五，建立风险评级机制。亦即，将风险进行分类评级，并根据各个风险等级制定不同层次的风险规制措施。例如，根据本文对深度伪造技术法律风险的类型化分析，对涉及风险类型 II（基于市场经济秩序的中观法律风险）、风险类型 III（基于国家公共安全的宏观法律风险）等风险等级高的伪造情形，服务提供者采取阻止措施的义务较重，须履行严格的审查义务；而对于风险类型 I（基于个体权利谱系的微观法律风险）等风险等级较低的违法和不良内容，服务提供者采取阻止措施的义务相对较轻，可以豁免一般性审查义务，但仍需尽到注意义务。

2. 风险控制阶段的法律优化路径

在风险控制阶段，明确立法规制模式的选择，构建层次分明的规范体系，更新法律责任的落实方案等，有助于在现实中化解深度伪造技术引发的法律风险。

第一，厘清不同立法模式之间的“破立关系”。深度伪造虽是一项破坏性技术创新，但其引发法律风险的作用逻辑并没有超出现行立法所调整的范畴，其不当运用仅仅是法益侵害手段的革新和法益侵害程度的加深，而并未改变侵犯法益的实质。[16] 从实践来看，迄今为止我国已出台的各级各类规范可以为约束深度伪造技术提供基础法律依据①，因而我国当前不必急于针对深度伪造技术进行专门立法。

第二，理顺各类法律规范之间的“纵横关系”。由于我国目前有关深度伪造技术的规范较为冗杂且体系化不足，故仍需构建层次分明的法律规范体系。具体而言，一是根据不同法律位阶处理好各类规范之间的纵向关系。例如，将现行有关约束深度伪造技术的各类规范按照“法律—行政法规—地方性法规—行政规章—其他规范性文件”的立法层级排列，同时处理好作为一般法的《民法典》《刑法》等与作为特别法的《网络安全法》《数据安全法》《个人信息保护法》等规范之间的关系。二是处理好不同部门颁布的同一位阶规范之间的横向关系，及时清理同位阶规范之间的冲突规则，同时提倡以多部门联合发文形式代替各部门单独发文。

第三，明确法律责任落实方法的“新旧关系”。一是坚持传统的责任落实方案，通过各部门法实现责任追究。事实上，无论利用深度伪造技术侵犯公民个体权利，还是破坏市场经济秩序，抑或危害国家公共安全，均可适用传统的责任落实方案，在事后对其予以行政处罚或提起诉讼，使其承担民事、行政或刑事责任。二是更新责任落实方案，尤其注重对信息规制工具的使用。例如，建立标识信用评级机制，对用户使用深度合成服务的标识情况进行信用评级，并建立“失信”档案。又如，引入“污染者自担”的信息更正制度，由技术滥用者或信息污染者发布更正声明，澄清误解。

（三）算法之治：以技术监督技术

由于深度伪造技术的生成机制具有高度不确定性，服务提供者难以控制大模型所生成的数据，故可以“技术监督技术”的方式来规范深度伪造的内容，即以代码表达规制，用计算机语言“转写”法律规则，从而实现规范的技术化约束。

首先，在深度合成前的算法规范层面，一是在深度伪造技术的取材过程中，服务提供者应依照《网络安全法》《数据安全法》和《个人信息保护法》等规定依法处理个人信息，训练大模型所使用的个人信息必须经过匿名化处理才能使

①传统的部门立法，如《民法典》《刑法》等；新兴的专门立法，如《网络安全法》《数据安全法》《个人信息保护法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《人脸识别技术应用安全管理规定》《生成式人工智能服务管理暂行办法》等。

用。二是在大模型预训练阶段，设计能够自动审核明显违法内容的算法并对其进行优化训练，如屏蔽敏感词、违法图像等素材，防止其生成违法内容。

其次，在深度合成后的算法规范层面，一是添加防伪标识，建立数字内容可信体系。任何经深度伪造的内容都可以通过区块链技术来实现实时记录和追踪溯源，确保深度伪造的内容更加透明和有据可循。例如，借助区块链技术为原始的视频、音频、图片等素材添加时间戳，在分布式账本上生成不可变的元数据记录，使内容从制作到传播的全过程都可以实时追踪溯源。[17] 二是强化反制措施，研发算法模型，增加数字内容检测的技术供给。例如，通过识别伪造的人脸和图像帧的剩余部分是否协调，或检测生成的内容是否缺少生理信号等，来辨别伪造的人脸等。

作为人工智能领域一项重要技术创新，深度伪造技术被广泛应用于社会各个领域，释放了巨大的应用价值。然而，任何技术都是一把“双刃剑”，深度伪造技术的应用在释放巨大价值的同时也给个体、市场和国家带来了法律风险。为寻求技术创新与风险防控之间的平衡，世界各国（地）都在积极地探索不同的治理路径。当前，我国应当充分吸收借鉴世界各国先进经验并结合我国实际，综合运用软法、硬法、算法等手段，对深度伪造技术可能产生的法律风险进行系统、有效防控。

[参考文献]

[1][澳]彼得·德霍斯.知识财产法哲学[M].北京:商务印书馆,2008:141.

[2][美]约瑟夫·熊彼特.经济发展理论——对利润、资本、信贷、利息和经济周期的探究[M].北京:中国社会科学出版社,2009:85.

[3][美]克莱顿·克里斯汀森,[加]迈克尔·雷纳.创新者的解答[M].北京:中信出版社,2013:25-26.

[4]Stefan Hüsig, Christiane Hipp and Michael Dowling, “Analysing Disruptive Potential: the Case of Wireless Local Area Network and Mobile Communications Network Companies”, R&D Management, Vol.35, No.1, 2005, pp.17-35.

[5]Andreas Keller and Stefan Hüsig, “Ex ante Identification of Disruptive Innovations in the Software Industry Applied to Web Applications: the case of Microsoft’s vs. Google’s Office Applications”, Technological forecasting & social change, Vol.76, No.8, 2009, pp.1044-1054.

[6]S.Y Lee et al., “Recognition of Human Front Faces Using Knowledge-based Feature Extraction and Neurofuzzy Algorithm”, Pattern Recognition, Vol.29, No.11, 1996, pp.1863-1876.

[7]Mika Westerlund, “the Emergence of Deepfake Technology: A Review”, Technology Innovation Management Review, Vol.9, No.11, 2019, pp.39-52.

[8]Mary Anne Franks and Ari Ezra Waldman, “Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions”, Maryland Law Review, Vol.78, No.4, 2019, pp.892-898.

[9]刘国柱.深度伪造与国家安全:基于总体国家安全观的视角[J].国际安全研究,2022(3).

[10]Jessica Newman, “Decision Points in AI Governance”, https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf,2023-12-20.

[11]Jonathan M. Gaffney, “Watching the Watchers: A Comparison of Privacy Bills in the 116th Congress”, https://crsreports.congress.gov/product/pdf/lsb/lsb10441, 2023-12-20.

[12]Daniel Castro, Michael Mclaughlin and Eline Chivot, “Who Is Winning the AI Race: China, the EU or the United States?”, https://www2.datainnovation.org/2021-china-eu-us-ai.pdf, 20-12-2023.

[13]European Political Strategy Center, "The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines", March 2018.
[14]Andreas Klinke and Ortwin Renn, "Expertise and Experience: A Deliberative System of A Functional Division of Labor for Post-normal Risk Governance", Innovation: The European Journal of Social Science Research, Vol.27, No.4, 2014, pp.442-465.
[15]赵鹏.私人审查的界限——论网络交易平台对用户内容的行政责任[J].清华法学,2016(6).
[16]姜瀛.人工智能深度伪造技术风险刑法规制的向度与限度[J].南京社会科学,2021(9).
[17]Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics", Foreign affairs, Vol.98, No.1, 2019, pp.147-155.

责任编辑：林珊珊

Research on Legal Risk Prevention and Control Countermeasures of Deepfake

Long Jun & Wang Tianyu

[**Abstract**] The Deepfake has realized the technological revolution from continuous innovation to destructive innovation, and has produced both positive and negative effects in its application. From a typological point of view, the Deepfake simultaneously triggers micro legal risk based on individual rights spectrum (type I), medium legal risk based on market economic order (Type II) and macro legal risk based on national public security (type III), and thus forms a diversified and hierarchical legal risk pattern. In order to effectively resolve risks, the United States, the European Union and China have respectively formed the "bottom-up" decentralized governance model, the "top-down" vertical governance model and the "multi-participation" cooperative governance model, and each presents distinct advantages and disadvantages. In view of this, China can achieve effective prevention and control of the above-mentioned legal risks by forming a comprehensive governance model combining "soft law + hard law + algorithm": at the level of soft law, we should perfect the system of science and technology ethics and clarify the substantive and procedural norms of science and technology ethics; at the level of hard law, the rigid constraints of risk prevention and control could be strengthened through the coordination of the two subsystems of risk prevention and risk control; at the algorithm level, we should realize the whole process technical supervision from sample data collection to deepfake content release, and give full play to the normative function of technology.

[**Key Words**] Deepfake; Disruptive Innovation; Legal Risk; Ethics of Science and Technology; Algorithm