Open Journal of Legal Science 法学, 2026, 14(3), 90-96
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2026.143071

Hans汉斯

深度伪造的技术逻辑、风险类型化与规制进路

成沁滢

宁波大学马克思主义学院，浙江 宁波

收稿日期：2026年2月1日；录用日期：2026年2月11日；发布日期：2026年3月13日

摘　要

从生成式对抗网络(GANs)到最新的扩散模型，深度伪造(Deepfakes)技术推动了视听资料从传统的“机械记录”向“智能生成”范式的根本性跃迁。这种技术红利在重塑数字内容生产方式的同时，也对既有的法律规制体系形成了一定冲击。它呈现出显著的自动化与低门槛特征，致使传统法律在认定侵权主体与归责原则时陷入两难。其风险不仅体现为对个人信息权益、知识产权等合法权益的侵蚀，更在宏观层面引发了公共信任危机，甚至影响了司法证据的真实性。鉴于此，单一的法律规制已显乏力，亟需构建一套整合技术防御、法律规范与社会协同的全链路治理框架，以期在释放技术潜能与维护法律价值之间寻求动态平衡。

关键词

深度伪造，生成式人工智能，算法安全，平台责任，全链路治理

Technical Logic, Risk Typology, and Paths to Regulation of Deepfakes

Qinying Cheng

School of Marxism, Ningbo University, Ningbo Zhejiang

Received: February 1, 2026; accepted: February 11, 2026; published: March 13, 2026

Abstract

Propelled by the evolution from Generative Adversarial Networks (GANs) to the latest diffusion models, Deepfake technology has catalyzed a fundamental paradigm shift in audiovisual materials, moving from traditional “mechanical recording” to “intelligent generation”. While this technological dividend reshapes the modes of digital content production, it also poses a structural challenge to existing legal regulatory frameworks. Characterized by high automation and low barriers to entry, it creates a dilemma for traditional legal mechanisms regarding the identification of infringing

文章引用: 成沁滢. 深度伪造的技术逻辑、风险类型化与规制进路[J]. 法学, 2026, 14(3): 90-96.
DOI: 10.12677/ojls.2026.143071

parties and the attribution of liability. The associated risks extend beyond the erosion of legitimate rights—such as personal information interests and intellectual property—to trigger a crisis of public trust at the macro level, even compromising the authenticity of judicial evidence. In light of this, reliance solely on legal regulation has proven insufficient. There is an urgent need to construct a "full-chain" governance framework that integrates technical defense, legal norms, and social coordination, aiming to seek a dynamic balance between unleashing technological potential and upholding legal values.

Keywords

Deepfakes, Generative AI, Algorithmic Security, Platform Liability, Full-Chain Governance

1. 引言

当人工智能从判别式迈向生成式(AIGC)新阶段，数字内容的生产逻辑正经历剧变。深度伪造技术作为其典型代表，依托深度学习算法，能自动解析海量样本特征，合成出肉眼无法辨识真伪的视听素材。这种伪造能力的质变，直接动摇了人类社会长期形成的"眼见为实"认识论。技术的泛化不仅引发了技术伦理的争议，更对现行法律体系构成挑战：如何确保人工智能在法治轨道上发展是必须回应的时代之问。

2. 深度伪造的技术逻辑与样态特征

深度伪造之所以成为法律规制的难题，根源在于其独特的技术逻辑。理解生成式对抗网络及扩散模型的运行机制，是准确界定其法律属性与责任边界的前提。

(一) 技术原理：生成式对抗与扩散模型的跃迁

深度伪造的核心技术基础是生成式对抗网络，包含两个核心神经网络：生成器与判别器。生成器的任务是利用随机噪声制造尽可能逼真地伪造数据，而判别器的任务则是区分数据的真伪。两者在训练过程中进行零和博弈：生成器不断优化算法以"欺骗"判别器，而判别器则不断提升鉴别能力，直至生成的内容达到以假乱真的程度[1]。

随着时代的发展，其技术架构已向扩散模型演进。这是一种基于概率生成的模型，通过正向加噪破坏数据结构，再通过反向去噪过程恢复数据，从而生成高质量图像[2]。这种"去噪复原"的生成逻辑，使得伪造内容在细节纹理上更加逼真，且能够实现跨模态生成，如文生视频。2024 年以来，深度伪造突破了非实时的瓶颈。现有研究中，攻击者利用流式扩散模型，如 StreamV2V，引入特征库机制，实现了毫秒级的推理延迟[3]。现如今，该类技术已被用于高度仿真的实时诈骗中：攻击者能在视频通话中模拟特定对象的外貌、声音与举止，实施难以辨识的欺诈，严重破坏远程身份验证与社会信任。

(二) 样态特征：不确定性、低门槛化与风险层级化

深度伪造技术呈现出不确定性与去中心化的特征，这从根本上异化了传统的侵权归责逻辑。传统数字侵权中，输入指令与输出结果的一一对应使得侵权具有确定性。而当前主流的扩散模型采取的是基于概率分布的随机生成机制。换言之，算法生成内容的过程并非对指令的机械执行，而是在潜在空间中对高斯噪声进行逐步去噪的随机采样过程。这种技术机理对法律认定构成了双重挑战：其一，模型的输出

结果具有不可预见性。即便开发者知晓用户的输入指令，也无法在事前预见模型最终生成的结果。其二，在因果关系判定上，概率生成机制在技术源头与侵权后果之间插入了随机性这一变量，切断了行为——结果的逻辑链条。这种技术性中断导致受害者难以在传统因果关系理论下，直接将特定的一张伪造图片追溯为开发者的直接加害行为。

近年来，深度伪造工具正以前所未有的速度完成“去专业化”进程，涌现出各类轻量化应用与在线平台。开发者将复杂算法封装进用户友好的移动应用或云端服务界面，用户无需具备高深的编程知识、高端硬件设施，只需上传照片、选择模板，即可在几分钟内获得伪造内容。这种一键生成的交互，剥离了技术背后的复杂性，使深度伪造从一门专业技能，转变为人人可操作的功能。这种技术平权化导致了风险源头的泛化，使得侵权主体从专业的黑客群体扩散至普通公众，极大地增加了法律规制和执法的成本。

除此之外，有学者对深度伪造样态特征进行了全面精准的概括：在技术本体层面，深度伪造表现出超拟真、反鉴别、快更迭与通用性强的四大特征，导致了算法黑箱与数据偏见等内生隐患；而在社会后果层面，风险呈现出从微观个体权益侵害，向中观市场秩序破坏，再到宏观国家安全威胁扩散的层级化特征[4]。

3. 深度伪造引发的法律风险类型化分析

深度伪造技术的广泛应用，使得法律风险从单一的私权侵害向公共安全领域蔓延，已深度嵌入个人信息权益、著作权、金融安全、司法证据效力及国家安全等多元法益场域。

(一) 肖像权、声音权和个人信息权益

深度伪造技术能够对目标人物的面部、声纹及体态进行高精度的合成与替换，制造出逼真却虚假的视听内容。这一过程触及了公民的生物特征信息，侵害肖像权、声音权等标识性人格权。深度伪造的运作逻辑依赖于对海量人脸、声纹等敏感生物识别数据的深度抓取与分析。即便某些伪造行为未在公共领域造成广泛传播，但其在算法训练阶段未经授权抓取、分析生物特征数据的行为，本质上已经构成了对个人敏感数据权益的侵害，破坏了主体对私密信息的安全边界与决断权。除此之外，这种技术强行割裂了自然人与其社会形象的真实纽带。当面部、声音等生物特征被非法挪用，不仅意味着个人对自我形象控制力的丧失，更导致“数字身份”被盗用或伪造。在数字化时代，这种身份层面的混同与篡改，极易引发外界对主体真实意愿和行为的误认。程啸教授认为，对此类风险应进行场景化区分：若行为主要表现为利用伪造影像进行商业宣传或侮辱诽谤，侧重于对肖像与名誉利益的公开利用；若行为主要表现为未经同意大规模收集、泄露人脸数据，则应重点规制其违反《个人信息保护法》关于敏感个人信息的处理规则的行为[5]。

(二) 名誉权和人格尊严

深度伪造技术常被用于制作侮辱性或色情内容，即便在物理上未接触受害人，也严重贬损了其名誉权和人格尊严，甚至可能引发网络暴力。深度伪造通过制造本人未曾有过的言行、表情或情境的虚假内容，强行割裂了真实人格与数字形象之间的统一性，戕害他人人格尊严。通过生成足以以假乱真的负面行为证据，它能够更精准、更高效地破坏受害者在特定职业圈、社交圈或公众中的声誉。更为严峻的是，数字内容的瞬时全球传播性与不可控性，使得受害者面临公开的、大规模的数字性羞辱，完全丧失对自身尊严处境的控制，极易导致社会性死亡及伴随而来的严重焦虑、抑郁等精神创伤，其损害范围与修复难度极大。

(三) 著作权和财产权

深度伪造技术同样对市场经济秩序构成威胁。随着刷脸支付的普及，人脸识别已成为金融账户安全

的关键防线。深度伪造技术通过生成具有活体特征的伪造人脸，可能突破现有的生物识别验证系统。而生物特征数据的不可更改性使得一旦泄露或被伪造，其后果将是灾难性的。一旦深度伪造技术被用于大规模金融诈骗，将不仅造成个人财产损失，更可能引发系统性的金融安全风险。而对于依赖个人形象与声誉获取收入的公众人物，其肖像和声音被滥用于低质或不当的商业推广，会迅速稀释其品牌价值与市场号召力。

在著作权和知识产权上，深度伪造引发了复杂的权属纠纷与侵权风险。深度合成的生成过程，本质上依赖对大量既有视听作品进行特征学习和内容重组，这一技术路径本身就埋下了版权风险。在模型训练环节，系统通常需要输入海量影视片段、音乐作品等数据，其中相当一部分并未取得权利人授权，因而可能已经触及复制权和信息网络传播权的边界；在生成环节，算法往往对原有作品的风格、形象甚至具体表达进行改写或变形，这种处理方式有时会损害作品完整性，也容易侵犯著作权和改编权。原作者的创作成果被拆解为数据特征，再以全新内容的形式出现，表面上看似与原作无关，实际上却可能建立在他人智力劳动之上，并进一步被商业化利用。

(四) 公共安全与社会信任

深度伪造技术正在改变人们理解信息真实性的方式，并进一步冲击社会信任的形成基础。有研究指出，生成式人工智能模型存在固有的“幻觉问题”，即模型生成的视听内容虽然在逻辑和感官上极具说服力，但往往与客观事实相悖。这种“一本正经地胡说八道”的特性，使得虚假信息披上了高可信度的外衣，导致公众难以通过传统的感官经验辨别真伪，从而严重削弱了数字内容的可信性基础[6]。深度伪造的常态化，容易将公众推向两种非此即彼的消极状态：一是陷入轻信，尤其容易被情绪化、符合自身偏见的伪造内容所俘获；二是陷入普遍的怀疑主义，对任何来源的信息甚至权威信源都预设不信任。前者制造盲动与恐慌，后者则导致共识的瓦解与合作的停滞，导致社会成员之间信任资本被透支，社会粘性减弱。

在司法审判、公共安全以及重大公共政策讨论等高度依赖事实判断的领域，深度伪造技术正在逐步侵蚀既有的信任基础。问题并不只在于个别虚假内容的出现，而在于由此引发的普遍性怀疑正在改变人们对证据和信息的基本态度。这种氛围一旦形成，社会运行所依赖的信任机制就会变得更加脆弱，相应的制度运行成本也随之上升，司法公信力和公共秩序都可能受到连带影响。在司法场景中，视听资料原本被视为具有较强直观性的证据形式，但在深度伪造技术介入后，其真实性越来越容易受到质疑。如果司法体系缺乏便捷而可靠的技术鉴别手段，围绕证据真伪展开的程序性争议势必增多，司法资源将被大量消耗在鉴定、质证和反复核查之中，从而拉长案件处理周期，也增加当事人的维权成本。在公共治理层面，当重要公共议题被刻意制造和传播的虚假信息所干扰时，公众判断现实的基础会变得不稳定，进而影响政策讨论所依赖的共识前提。信息来源的碎片化与真假难辨，使得治理过程更容易受到情绪化舆论的牵引。在自然灾害或突发公共事件等关键时间窗口内，如果有人利用深度伪造手段定向投放误导性内容，不仅可能干扰应急决策节奏，还可能放大社会恐慌或对立情绪，使本就复杂的治理情境进一步恶化，并显著提高协调与治理成本。

4. 全链路社会协同治理范式的构建

面对深度伪造技术的复杂性，单一的法律规制恐怕难以奏效，需要构建涵盖技术防御、法律规制与社会协同的全链路治理范式。

(一) 技术治理：构建全生命周期的防御体系

针对深度伪造技术可能引发的系统性风险，技术治理层面有必要通过多种手段的协同应用来提升整体防护能力。

在源头治理上，必须确立“隐私优先”的计算范式，将防护机制前置于数据生命周期的起点。面对训练数据采集加工中潜藏的侵权风险，引入成熟的隐私计算技术显得尤为关键。“可用不可见”的理念能够确保算法在不通过暴露原始数据的前提下实现有效训练。现有研究表明，通过差分隐私机制对数据注入控制噪声，可以使得单条数据的加入或删除对整体模型输出几乎无影响，从而有效避免模型记忆单个敏感信息而被反推泄露隐私；而同态加密技术则允许在密文状态下执行必要的计算操作，保障整个计算过程的机密性[7]。这种双重防御体系，既能降低敏感数据泄露的直接风险，又能增强对诸如数据投毒等恶意输入的抵抗能力。

在模型应用与检测层面，需要构建能够适应伪造技术快速迭代的检测体系。单一模态的检测方法往往在面对新型合成样本时表现力不足，因此基于多模态特征融合的算法成为当前研究重点。有研究提出了一种基于多域特征融合的多分支网络框架 MBMD，综合利用频率域、空间域和时空域信息来挖掘伪造线索。通过捕捉图像细微结构变化的频率特征、局部异常区域的空间特征以及全局时空不一致性，该技术能够显著提升对未知伪造类型的泛化检测能力[8]。

在确权与溯源层面，应依托区块链与数字水印技术构建不可篡改的信任锚点。国家数据局印发的《可信数据空间发展行动计划(2024~2028 年)》明确指出，区块链与隐私计算技术是构建可信数据空间、促进数据合规高效流通的基础技术底座。其核心在于构建一个多方互信的数据流通环境：在存证方面，可将原始内容的数字指纹及其生成者、时间戳等存证信息，记录在分布式账本上，为事后司法鉴定和权责追溯提供具有法律效力的技术铁证；在授权与溯源方面，通过结合数字身份与区块链智能合约，可以实现精细化的数据授权访问和全过程追踪。

(二) 法律治理：完善义务体系与责任链条

厘清算法服务提供者与使用者的责任归属是法律治理的核心。法律应明确规定技术开发者(指研发核心算法、训练基础模型的主体)在算法设计阶段的注意义务，以及服务提供者(指调用基础模型接口或基于开源模型进行微调，面向终端用户提供具体应用的主体)在内容发布阶段的审核义务，防止技术被滥用于制造虚假信息或侵犯他人权益。

在司法实践中，如何认定深度伪造技术开发者的责任是核心难点。围绕归责原则与过错认定标准，学界尚未形成一致意见。有学者主张“无过错责任”，将生成式人工智能侵权视为一种具有高度风险的新型危险活动。该立场认为，技术开发者通常从相关技术服务中获得经济利益，也更有能力通过成本分摊机制分散风险；相比之下，受害人往往处于信息与技术能力的弱势地位，举证证明开发者存在过错存在现实困难。在此背景下，适用无过错责任有助于强化受害人救济[9]。但也有学者持相对宽容的态度，主张仍应以“过错责任”为基本立场。他认为如果对开发者施加过重的责任负担，可能抑制技术研发与产业创新的积极性，从而对整体技术进步产生不利影响[10]。在严格责任与一般过错责任之间，有学者提出以“现有技术水平”作为衡量开发者是否尽到合理注意义务的客观标准，主张“过错推定”原则[11]。这一思路强调，“现有技术水平”并非静态指标，而是随时间、行业发展状况及技术能力变化而动态调整。如果开发者能够证明，其在相关技术条件下已采取符合行业通常标准的风险防范措施，但损害仍然难以避免，则不宜轻易认定其存在过错。具体而言，可从输入端与输出端两个环节进行审查，例如是否对训练数据进行必要的合法性审核与去标识化处理，是否部署了与当时技术水平相适应的内容审核与合成标识机制[11]。在平衡技术创新与风险防控的现实语境下，“过错推定”原则是更具合理性与可行性的制度选择。

随着深度合成技术与算法推荐机制的深度融合，平台已从单纯的信息存储空间演变为内容生态的主动组织者。这种角色的转变使得平台对深度伪造内容的传播具有了更强的控制力与获益性，从而引发了

其法律注意义务的实质性升级。确定平台及服务提供者的法律责任，关键在于厘清技术赋能下的权利义务边界：即在承认技术中立的同时，必须强化其对高风险内容的注意义务。为此，法律应构建一套动态的责任认定机制，依据平台对“强制标识”“算法审核”及“应急处置”等法定义务的履行情况，精准划分三类责任：对于未落实显著标识与隐性水印等法定技术规范导致内容无法溯源的情形，系直接违反《深度合成规定》第十六条、第十七条设定的“内容标识义务”，属于平台自身法定义务的不履行，应由平台承担直接责任；对于算法未能识别显而易见伪造内容仍进行置顶或精准分发的情形，依据《民法典》第一千一百九十七条关于网络服务提供者“知道或者应当知道”规定，可认定平台在算法推荐中未尽到合理的注意义务，存在主观过错，需与直接侵权人承担连带责任；对于因未落实实名制而导致直接侵权人无法锁定的情形，鉴于平台违反了《网络安全法》关于用户身份管理的前置性审核义务，导致受害人索赔对象缺失，应参照《民法典》第一千一百九十八条“安全保障义务人责任”规定，判定平台在防止损害扩大的范围内承担补充责任。通过这种分层归责体系，旨在实现技术创新与社会安全的动态平衡。

对于技术开发者与服务提供者，重点在于强化其事前预防义务，即通过算法备案、强制标识等技术手段，防范风险的源头生成；而对于技术使用者，则应重点规制其事后行为后果。首先，必须明确严守权利边界与维护标识完整性的核心义务。除法律规定的合理使用情形外，严禁使用者在未获明确授权的前提下，不得利用深度合成技术处理他人肖像、声音等人格要素；同时，使用者在发布时负有主动声明与显著标识的法定义务，任何恶意利用技术手段篡改、隐匿平台预置隐性水印的行为，均应被认定为破坏溯源机制的主观恶意，作为认定侵权或行政处罚的从重情节。其次，应构建“民行刑”梯次配置的追责体系以实现有效威慑。在民事领域，依据《民法典》确立不以“营利为目的”的侵权认定标准，畅通停止侵害与损害赔偿的救济渠道；在行政领域，对利用深度伪造制造谣言、扰乱公共秩序但尚未构成犯罪的行为，由网信与公安部门实施罚款或账号禁用的快速惩戒；在刑事领域，则对利用该技术实施精准诈骗、侮辱诽谤或危害国家安全等严重犯罪活动进行严厉打击。

(三) 社会协同：共治格局的形成

治理深度伪造是一场对技术伦理、社会信任和公众认知的综合考验，需要构建一个政府主导、行业担责、公众参与、社会监督的多元协同共治格局，形成社会合力。

在行业层面，可通过强化自律机制提升源头治理能力。头部科技企业作为技术源头，应联合组建行业联盟，制定通用的深度合成内容标识协议与元数据标识，有助于提高跨平台治理的协调性。同时，通过风险信息共享和违规应用通报等方式，也能在一定程度上压缩灰色应用的空间；在社会层面，应为第三方核查力量的发展提供条件。高校、研究机构及专业组织在技术评估方面具有一定优势，可以为媒体和公众提供相对独立的鉴别支持。与此同时，新闻媒体在引用网络素材时加强来源核实，也有助于减少失实信息在主流渠道中的扩散；在公众层面，应将应对深度伪造纳入全民数字素养教育体系，普及批判性思维与识谣防骗能力，如技术原理、常见破绽以及可信信息验证的基本方法，以降低普通用户在复杂信息环境中的误判风险。

5. 结语

深度伪造技术的迭代与普及，标志着人类社会进入了后真相时代的数字深水区。技术不仅重塑了信息的生产方式，更深刻地介入了法律所保护的社会关系。法律的滞后性与技术的敏捷性之间的张力将长期存在，但这正是法治进化的动力。面对这一挑战，法律规制不能仅局限于对具体侵权行为的末端修补，而应向技术源头回溯，重新厘定算法时代的权利义务边界。唯有构建多元主体参与的治理体系，才能有效化解技术快速发展引发的规制难题，确保技术创新在法治框架内运行，最终达成安全保障与产业发展的动态平衡。

参考文献

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. and Bengio, Y. (2014) Generative Adversarial Nets. *Communications of the ACM*, **63**, 139-144.

[2] Ho, J., Jain, A. and Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, **33**, 6840-6851.

[3] Liang, F., Kodaira, A., Xu, C., *et al*. (2025) Looking Backward: Streaming Video-to-Video Translation with Feature Banks. *Proceedings of the International Conference on Learning Representations* (*ICLR*), Singapore, 24-28 April 2025.

[4] 黄静, 韩松言, 田宇航. 生成式人工智能深度伪造风险的样态特征、生成逻辑与监管策略[J]. 电子政务, 2025(5): 31-41.

[5] 程啸. 个人信息权益与标识性人格权的关系[J]. 国家检察官学院学报, 2025(3): 41-57.

[6] 刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025, 36(3): 1152-1185.

[7] 陈全涛, 张仰森, 王璞, 等. 生成式人工智能安全风险与防御策略综述[J/OL]. 计算机科学, 1-23. https://link.cnki.net/urlid/50.1075.TP.20251017.1447.016, 2026-01-27.

[8] 龙敏, 尹茜, 张乐冰, 等. 基于多域特征融合的多分支网络用于 Deepfake 检测[J]. 中国图象图形学报, 2026, 31(1): 120-137.

[9] 徐伟. 生成式人工智能服务提供者侵权归责原则之辨[J]. 法制与社会发展, 2024(3): 190-202.

[10] 王利明. 再论生成式人工智能的侵权风险及其应对[J]. 广东社会科学, 2026(1): 247-252.

[11] 王若冰. 论生成式人工智能侵权中服务提供者过错的认定——以“现有技术水平”为标准[J]. 比较法研究, 2023(5): 20-29.