国家安全视阈下深度伪造的社会安全风险及应对

郑豪男

西南政法大学，重庆市，401120；

摘要：[研究目的] 国家安全视阈下，深度伪造技术的迅猛发展正突破传统信息防伪体系的防御阈值，催生出系统性社会安全风险。深入剖析深度伪造技术引发的社会安全风险生成机理、具体表现及应对策略，可为维护国家社会稳定与公共安全提供理论支撑与实践指导。[研究方法] 通过文献分析、案例研究、比较研究、跨学科分析等方法，围绕社会安全的概念内涵、深度伪造风险传导逻辑、多维风险表达及综合治理路径开展研究。[研究结果 / 结论] 社会安全是涵盖社会秩序稳定、公共信任存续、制度权威有效运行等要素的安全范畴，深度伪造生成社会安全风险遵循 “技术滥用 — 信任崩塌 — 秩序失序” 的三阶逻辑。其风险具体表现为消解公共信任根基、扰乱社会秩序稳定、侵蚀制度权威合法性。对此，构建技术防御体系、法律规制框架、公众认知教育、国际协同治理四维应对体系，可系统性化解深度伪造时代的社会安全风险。

关键词：深度伪造技术；社会安全；国家安全；技术治理；数字信任

DOI：10.64216/3080-1486.25.04.052

引言

随着生成式人工智能技术的迭代演进，深度伪造技术通过生成式对抗网络、扩散模型等算法机制，能够以低成本、高效率生成高度逼真的虚假音视频内容，实现对人物言行的精准模拟与篡改。这一技术在影视后期制作、虚拟主播、历史影像修复等领域展现出创新价值，但其“以假乱真”的内容生产能力已突破传统信息传播的真实性边界，促使信息生态从“真伪可辨”向“虚实难分”范式跃迁，标志着人类社会进入“算法介导型虚假信息泛滥”的新阶段。这种技术赋能的虚假信息生产，不仅重塑社会信息传播的技术路径，更通过对公共信任体系的精准冲击，持续消解主流社会价值共识，挤压国家社会安全空间，催生出系统性社会安全风险。

在国家安全视阈下，深度伪造技术的风险传导已超越单纯的技术风险范畴，演变为影响社会稳定、挑战治理能力的非传统安全威胁。深入解析其风险生成逻辑，识别风险具体表现，并构建多维度应对体系，对于维护国家社会安全具有重要的理论与实践意义。

1 相关研究

通过对现有社会安全相关研究梳理可知，人工智能技术的不断迭代更新给社会安全带来的更多的挑战和问题。国内方面的研究从国家安全角度出发，主要是围绕深度伪造对于社会的威胁的路径研究，学者李柯指出，深度伪造在虚假信息传播、政治认同削弱、治理行动干扰三方面的危害路径，提出“技术－观念－制度”三维治理框架。郭恩泽从政治、经济、社会安全多维度剖析深度伪造的威胁，主张以“技术研发+立法监管+国际协作”构建安全共同体。赵建强指出深度伪造对司法证据效力的颠覆性影响，建议通过数字水印、区块链构建可信内容溯源体系。邹卫中揭示深度伪造对个人、组织、国家的“工具性－观念性－关系性”复合风险，批判现行“包容性监管”的不足，主张市场与公权力协同治理。国内学者对于社会安全的理解与国家安全紧密合，刘国柱认为深度伪造不仅给国际形势和国际安全环境带来冲击，恶化“信任赤字”，还在国防、政治、经济和国民安全等领域制造复杂问题。以上这些观点也共同强调了国家安全视阈下技术因素在社会安全的重要性以及应对相关风险的必要性。

国际方面对于深度伪造下的社会安全研究主要在于监管与预防，Citron,D 提出“针对性伤害规制”理论，主张以民事赔偿优先原则保护深度伪造受害者。Alanazi, S. et al.基于法学家、AI 伦理学家访谈，设计“技术-教育-法律”三维治理模型。国外学者的研究为我国提供了一定的帮助，展示了新的思路的同时，又在实际案例应用中以国家安全为导向防范技术驱动的社会安全风险提供了一定的理论和实践参看意义。

既有研究虽从不同维度揭示了深度伪造的风险特

征与治理方向，但缺乏从国家安全视阈对社会安全风险的系统性解构，尤其在风险生成的内在逻辑、多维度风险的传导机制及国家层面的战略应对体系方面仍需深化。本文立足总体国家安全观，构建深度伪造社会安全风险的分析框架，为治理实践提供理论支撑。

2 社会安全内涵的界定

“社会安全”作为国家安全体系的核心要素，其概念内涵随技术革命与社会变迁持续拓展，在深度伪造技术背景下呈现出新的时代特征。

结合《国家安全法》对“国家安全”的“状态能力说”定义，国家安全视阈下的“社会安全”可界定为：“国家通过制度安排与技术手段，确保社会信息生态真实可控、公共信任体系稳定有序、价值共识免受内外干扰的状态，以及保障这种状态持续存在的能力。”这一概念既继承总体国家安全观“统筹发展与安全”的核心思想，又回应了深度伪造技术带来的新挑战，为风险识别与治理提供了理论坐标。

3 深度伪造生成社会安全风险的内在逻辑

深度伪造的社会安全风险生成遵循“技术滥用—信任崩塌—秩序失序”的三阶逻辑，其风险传导呈现螺旋式升级特征，各环节相互强化，形成系统性风险链条。

3.1 技术滥用：风险生成的起点

深度伪造技术的“去中心化”应用降低了虚假信息生产门槛，使个人与非国家行为体可低成本制造具有高度迷惑性的虚假内容，构成风险生成的初始动力。

国家或非国家行为体将深度伪造作为认知战工具，用于政治操纵、社会煽动等系统性破坏活动。外国势力通过伪造我国领导人不当言论视频制造外交危机，极端组织利用虚假灾难新闻煽动社会对立，反政府势力通过合成军警暴力执法视频抹黑政府形象等。这类滥用具有“规模效应”与“政治意图”，直接冲击社会稳定与国家政权安全。

3.2 信任崩塌：风险扩散的关键环节

社会信任体系依赖“信息真实性”这一认知基础，深度伪造通过模糊“真实”与“虚假”的边界，使公众陷入“认知不确定性”困境，构成风险扩散的关键传导节点。其信任破坏机制体现在对信息源信任的侵蚀。传统社会中，公众通过“信源权威性”判断信息真实性，而深度伪造技术可精准模仿权威信源的内容风格与呈现形式，使公众难以区分真伪。伪造央视新闻播报虚假政策，合成专家访谈传播错误健康知识，这类行为直接削弱权威信源的公信力，导致“塔西佗陷阱”——当政府或权威机构的信息真实性受到普遍怀疑时，即便发布真实信息也难以获得信任。

这种信任崩塌具有累积效应，初期可能表现为局部性信任受损，但随虚假信息持续扩散，会逐渐侵蚀社会信任的整体根基，为风险升级奠定心理基础。

3.3 秩序失序：风险激化的最终表现

社会秩序的维系依赖共识性认知与制度权威，深度伪造通过以下路径解构秩序，构成风险链条的最终环节。煽动社会对立，破坏社会整合。干扰政策执行，削弱治理效能。动摇司法公正，挑战法治权威。

当这些风险突破社会承受阈值，可能引发系统性动荡，威胁国家安全。这种从技术滥用开始，经信任崩塌传导，最终导致秩序失序的逻辑链条，构成了深度伪造社会安全风险的完整生成机理。

4 深度伪造引发的社会安全风险表达

基于上述内在逻辑，深度伪造引发的社会安全风险在具体场景中呈现出多维表达，可从微观、中观、宏观三个层面进行解构。

4.1 微观层面：个人权利侵害与社会恐慌

深度伪造对个体权益的侵害已形成规模化风险，直接影响公众安全感，构成社会安全的微观威胁。 深度伪造技术已成为电信诈骗的“新工具”。这类案件的成功率远高于传统电信诈骗，因其利用了熟人关系的信任基础，使受害者在短时间内难以辨别真伪。

4.2 中观层面：公共信任消解与社会分裂

深度伪造对中观层面社会系统的冲击，主要表现为公共信任体系的瓦解与社会群体的极化对立，削弱社会整合能力。

媒体公信力遭遇“断崖式下跌”。传统媒体作为“事实核查者”的角色被深度伪造技术解构，虚假新闻通过算法推荐快速扩散，而真相的澄清往往滞后。这种“虚假信息先行”的传播模式，使媒体的权威性持续受损，公众信息获取陷入“渠道焦虑”。

社会群体出现“认知隔离”与“对立极化”。深度

伪造技术通过精准捕捉不同群体的认知偏见，推送定制化虚假信息，加剧群体间的认知鸿沟。这种群体极化若得不到有效控制，可能演变为线下冲突，破坏社会和谐稳定。

4.3 宏观层面：制度权威弱化与国家安全挑战

深度伪造对宏观层面国家治理体系的冲击，主要表现为制度权威的合法性受损与国家安全防线的松动，构成战略性风险。

政治权威面临“合法性危机”。外国势力、反政府势力利用深度伪造技术，伪造国家领导人、政府官员的不当言论、决策失误等内容，误导公众对政府的认知，削弱政治信任。这类行为直接挑战政治权威，干扰国家治理进程。

国家安全防线出现“认知漏洞”。造军事演习视频泄露虚假机密，合成将领讲话误导国防决策；篡改国际会议录音引发外交争端，伪造他国领导人表态破坏国际关系；这些行为通过干扰认知判断，影响国家战略决策，对国家安全构成间接但深远的威胁。

5 国家安全视阈下深度伪造社会安全风险的应对策略

应对深度伪造的社会安全风险，需构建“技术防御—法律规制—认知赋能—国际协同”的四维体系，实现风险全链条治理，在保障技术创新的同时筑牢社会安全防线。

5.1 技术防御体系

技术防御是应对深度伪造风险的基础工程，需通过技术创新提升风险识别、溯源与处置能力，构建主动防御体系。

5.1.1 完善内容溯源机制

推广“数字水印”技术，要求所有生成式 AI 工具在输出内容中嵌入不可篡改的标识信息，明确内容的生成主体、生成时间、技术类型等元数据。抖音、快手等短视频平台可要求用户发布的AI生成内容必须附带“AI生成”水印；新闻机构使用 AI 技术制作的虚拟主播报道，需在画面角落持续显示“虚拟合成”标识。

建立“内容溯源区块链平台”，对音视频、图片等内容的创作、传播过程进行全程存证，确保内容可追溯。推行“深度伪造内容分级标识制度”，根据内容风险等级采用不同标识方式。对娱乐性质的 AI 换脸视频采用蓝色标识；对涉及政治、军事、医疗等敏感领域的 AI 生成内容采用红色标识，并附加风险提示；对无法确认真实性的内容采用黄色标识，提醒用户谨慎判断。

5.2 法律规制框架

法律规制是应对深度伪造风险的制度保障，需通过立法、执法、司法全链条完善，构建清晰的行为规范与责任体系。

5.2.1 完善法律法规体系

修订现有法律法规，填补深度伪造风险治理的法律空白。在《刑法》中增设“深度伪造罪”，对利用深度伪造技术实施危害国家安全、破坏社会秩序、侵害他人权益等行为设定刑事责任；在《民法典》中细化“个人信息保护”条款，明确未经同意使用他人肖像、声音进行深度伪造的民事侵权责任；在《网络安全法》中增加“算法透明度”要求，规范平台利用算法推荐深度伪造内容的行为。

5.2.2 健全执法司法机制

建立“深度伪造案件专项侦查队伍”，提升执法能力。在公安部、省级公安厅设立专门机构，配备“网络安全+刑事侦查+AI 技术”的复合型人才，负责重大深度伪造案件的侦破；开发深度伪造案件侦查工具，如溯源追踪系统、证据固定软件等，提升办案效率；与互联网企业建立数据共享机制，快速获取案件相关的用户数据、IP 地址等信息。

5.2.3 推动行业自律规范

制定《深度伪造行业自律公约》，引导企业履行社会责任。鼓励互联网企业、AI 技术公司、内容平台等签署公约，承诺不开发、不传播危害社会安全的深度伪造技术；建立企业内部伦理审查机制，对新研发的深度伪造技术进行伦理风险评估；定期发布社会责任报告，公开深度伪造内容治理成效。

6 结语

深度伪造技术的双刃剑效应，使其成为国家安全视阈下社会安全治理的新课题。其引发的个人权利侵害、公共信任崩塌、社会秩序失序等风险，本质是技术异化下的信息生态失衡。

构建全方位的深度伪造社会安全风险应对体系，既要通过技术创新提升风险识别与防御能力，又要通过制度建设明确行为边界与责任；既要通过公众教育增强社

会免疫力，又要通过国际合作破解跨境风险难题。这一体系的构建，旨在实现“技术可控、风险可防、秩序可保”的治理目标，在保障技术创新活力的同时，筑牢国家社会安全防线。

参考文献

[1]李柯.深度伪造对国家信息安全的影响及应对策略[J].信息安全与通信保密,2025,(05):67-75.

[2]李媛媛.深度伪造应用实践、应用风险及规制研究[J].网络安全技术与应用,2024,(03):121-123.

[3]郭恩泽.总体国家安全观下深度伪造危害国家安全的应对路径[C]//《法治实务》集刊 2023 年第 3 卷——国家安全工作研究文集.广西警察学院反恐怖与反邪教研究中心;,2023:29-35.DOI:10.26914/c.cnkihy.2023.056778.

[4]邹卫中,张力耘.深度伪造安全风险的社会特性与多元化部门协作治理[J].江汉大学学报(社会科学版),2023,40(03):32-42+125-126.DOI:10.16387/j.cnki.42-1867/c.2023.03.003.

[5]赵建强,张辉极,杜新胜,等.人工智能安全的隐忧:深度伪造技术的挑战与应对[J].中国安防,2021,(06):40-47.

[6]Chesney R , Citron D .Deepfakes and the New Disinformation War[J].Foreign affairs, 2019, 98(1):147-155.

[7]McGLYNN C ,TOPARLAK T R .The ‘new voyeurism’ : criminalizing the creation of ‘deepfake porn’ [J].Journal of Law and Society,2025,52(2):204-228.

[8]雒辛芃,闵浩楠.社会安全事件网络舆情事理图谱构建与预测应用研究[J].新闻传播,2025,(05):28-30.

[9]黄大慧.国家安全学基础理论[M].北京:时事出版社,2024:184.

[10]李文良.国家安全学基础理论框架构建研究[J].国际安全研究,2022,40(5):3-29,157.

作者简介:郑豪男，男（1998-04-23），汉族，浙江人，硕士，研究方向:国家安全。