公共安全视角下深度伪造风险的概念模型与生成机理

詹国彬　陈逸凡

摘　要　数字时代背景下，深度伪造技术如同魅影般在网络空间肆意游荡，凭借其以假乱真的特性对公共治理的真实性、透明度以及公信力构成了前所未有的挑战，同时也对公共安全乃至国家安全造成了重大威胁。研究发现，深度伪造风险的生成是一个多机制协同作用的动态过程：技术资本化是起点，驱动深度伪造技术的快速商业化应用；资本权力化助推深度伪造技术异化为社会操控工具，通过制造虚假信息影响公众认知；风险催化机制因技术低门槛和平台算法操控而加剧传播；正反馈机制导致虚假内容在传播中不断自我强化；最终借助风险传导机制形成跨领域、多层次的危害放大效应。为此，应从法律规制、政策监管、制度规范、国际合作、公众参与等维度入手，构建一个复合性的风险治理框架，以规范和推动深度伪造技术在公共治理领域的合法应用与有序发展。

关键词　深度伪造风险　风险治理　概念模型　生成机理

作者詹国彬，南京审计大学国家安全学院教授（江苏南京 211815）；陈逸凡，南京审计大学国家安全学院硕士研究生（江苏南京 211815）。

中图分类号 D0　　**文献标识码** A　　**文章编号** 0439-8041(2025)08-0082-13

在人类文明漫长的发展过程中，人们早已习惯于依赖自身的感官系统，如视觉和听觉，来构建对世界的认知。这种依赖逐渐形成了一种深植于潜意识的“心理图式”，并在现代信息社会中得以延续。相比于抽象的语言描述，人们更倾向于相信自己所看到和听到的内容。现代传媒技术正是利用这一心理倾向，试图通过更直观的视听媒介建立新的交流方式以增强可信度。例如，相较于单纯的文字描述，音频因其与真实语音的高度相似性，被视为更贴近现实的表达形式。因此，在现代信息社会中，“有图有真相”的固有认知逐渐被“绘声绘色”的图像、音频或视频所取代，后者因其更强烈的感官冲击力而被认为更具说服力。在部分技术乐观主义者看来，这似乎是信息技术推动社会进步的又一例证。然而，随着人工智能技术的飞速发展，特别是“深度伪造”（Deepfake）的出现，这一论证被蒙上了一层阴影。

“深度伪造风险”（Deepfake Risk）是指基于生成对抗网络中“生成器”（generator）与“鉴别器”（discriminator）两种神经模型的内部博弈[①]，所生成的虚假图像、音频和视频等内容，因其高度逼真性和易传播性对个人隐私、社会信任、经济秩序和政治安全等多领域造成的潜在危害。近年来，深度伪造技术以其逼真的场景再现能力，在文化、教育、数据安全等领域为公众生活带来了诸多便利，但同时所引发的治理问

① Robert Chesney & Danielle 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, 2018, pp. 882–891.

题也可谓不胜枚举。例如，2025 年 3 月 4 日上午，全国人大代表、小米集团创始人雷军发文，就其在十四届全国人大三次会议中提出的“重点治理 AI 换脸拟声治理重灾区”的建议进行说明和呼吁。[①] 此消息一经发布便登上了各大媒体的头版头条，引发了全社会对这一新型社会风险的广泛讨论。同年，3 月 7 日，国家互联网信息办公室、工业和信息化部、公安部、国家广播电视总局联合发布《人工智能生成合成内容标识办法》，明确要求对 AI 生成合成内容进行标识，并制定了一系列规范性措施。[②] 这一文件的出台预示着针对深度伪造风险治理的“政策之窗”可能即将开启。在此背景下，如何科学审慎地对待这一兼具创新潜力与风险特性的新兴技术，努力实现其正向价值最大化与负面效应最小化的动态平衡，对完善数字时代的技术治理体系与实践路径具有重要的学术价值与现实意义。

一、文献回顾与问题提出

当前，针对深度伪造风险及其治理的研究主要基于如下三个视角：第一，深度伪造风险的缘起。由数字技术所驱动的新一轮科技、产业革命对国际格局产生了巨大的冲击[③]，在这其中，因社交媒体的兴起导致虚假信息激增，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。[④] 一方面，市场需求的推动使技术门槛大大降低，“AI 换脸”等深度伪造的一键式操作成为可能[⑤]，并迅速席卷整个国际社会，使得针对技术风险的防治措施变得极为困难。[⑥] 另一方面，深度伪造技术发展过于迅猛，各类民事或刑事案件层出不穷[⑦]，现行法律的修订速度逐渐跟不上技术风险的发展与蔓延速度[⑧]，导致针对深度伪造风险的事后治理陷入无力的状态，常常如钟摆般处于“治理不利”或“过度治理”的矛盾中[⑨]，进一步引致风险的滋生。第二，深度伪造风险的样态。在国际政治领域，针对政治人物所进行的深度伪造攻击屡见不鲜[⑩]，例如，“在某些政治事件中，有人利用深度伪造技术制作虚假视频，声称某位领导人发表了不实言论或做出了不当行为，进而在国际社会引发混乱和误解”[⑪]。与此同时，深度伪造在经济领域中滥用所产生的风险也成为另一类亟需被加以重视的问题，例如，在经济和金融危机时期，深度伪造可能利用和放大先前存在的经济担忧和市场波动。[⑫] 抑或是模拟金融界的重要人物的影像和声音以便实施诈骗，以攫取不义之财。[⑬] 最后，在民生领域中深度伪造以一种“娱乐至死”的方式侵蚀和干扰着社会的正常秩序。深度伪造适配了互联网尤其是短视频“短平快”的平台特征以及“注意力经济”的平台逻辑[⑭]，助推公众热衷于获取短时间的刺激性快感，用深度伪造技术进行“恶搞”，导致公民个人的肖像权和隐私权被无限制侵犯。[⑮] 第三，深度伪造风险的防治策略。目前，针对深度伪造风险的防治策略有两种：其一，主动防御策略。该策略侧重防患于未然，强调在风险尚未形成时，便通过技术、法律、教育等手段，从源头上遏制深度伪造的滥用。例如，通过在 AI 影视作品中加入水印、建立

① 文丽娟：《代表委员建言加强 AI 虚假信息治理》，《法治日报》2025 年 3 月 6 日。
② 王思北：《强化全流程管理引导技术向善》，《新华每日电讯》2025 年 3 月 17 日。
③ 张力：《数字时代的国际秩序前景与中国方案》，《人民论坛·学术前沿》2024 年第 9 期。
④ 张立伟：《AI 技术可供性与媒体核心需求的适应性研究》，《当代传播》2025 年第 1 期。
⑤ 王新哲、杨建、马多贺：《面向 AIGC 对抗的人脸深度伪造检测方法研究》，《工业信息安全》2022 年第 11 期。
⑥ 吴进娥、冯树春：《“深度伪造”技术对法官认知的冲击及应对》，《学习与探索》2023 年第 9 期。
⑦ 李小波、郝泽一：《警察执法道德困境：一个可能的解释框架》，《北京联合大学学报（人文社会科学版）》2020 年第 4 期。
⑧ 王彦雨、李正风、高芳：《欧美人工智能治理模式比较研究》，《科学学研究》2024 年第 3 期。
⑨ 冯明昱、姜涛：《“深度伪造”滥用行为的刑法规制》，《湖北社会科学》2023 年第 4 期。
⑩ 刘俊、贾奕星：《肯定式生成与否定式观看：人工智能技术下影像创作与接受的一个悖论》，《未来传播》2025 年第 1 期。
⑪ 罗幸、黄鑫磊：《人工智能时代国际传播中的话语体系建构》，《传媒》2025 年第 4 期。
⑫ 刘国柱：《深度伪造与国家安全：基于总体国家安全观的视角》，《国际安全研究》2022 年第 3 期。
⑬ 刘晓丽、崔文波、张涛：《生成式人工智能视域下情报分析算法风险多重治理机制研究》，《图书与情报》2024 年第 5 期。
⑭ 陈思函、解学芳：《AIGC 驱动下的数字文化消费：困境透视与纾解路径》，《新疆社会科学》2024 年第 4 期。
⑮ 陆青：《数字时代的身份构建及其法律保障：以个人信息保护为中心的思考》，《法学研究》2021 年第 5 期。

内容认证机制等技术手段防止“以假乱真”的风险[①]，抑或通过推动立法，规范技术使用[②]，以及增强公民数字教育以提升其辨别能力等方式[③]，以防范和控制风险的滋生。其二，被动防御策略。该策略旨在深度伪造风险发生后通过检测、响应和补救措施，以减少其“涟漪效应”和负面影响。例如，采用国际合作的方式打击技术滥用分子[④]，采取行政命令紧急删除网上的深度伪造产品[⑤]，或在网络平台中嵌入对抗生成网络（GANs）检测工具，对虚假内容进行过滤。[⑥]

审视现有研究进展，大多数机构或学者针对深度伪造的“前世今生”开展了积极、广泛和有益的探讨，尤其是从技术应用层面提出了诸多富有价值的防治策略，已有研究为理解和考察深度伪造风险治理提供了富有洞见的参考，同时也为后续的研究奠定了基础和方向。但是，在深度伪造风险究竟是缘何而生的问题上，现有成果明显存在碎片化、分散化的缺憾，作为一种新型且极为复杂的现代社会风险，仅仅以“市场需求”或“技术推动”等角度作为研究切口加以讨论未免有失偏颇，缺乏整体性视角，研究的体系化和研究深度均有待加强。尽管少量研究针对深度伪造风险的演化过程进行了实证分析，但多为基于单一案例的事后复盘或是经验总结，其研究结论的普适性和解释力有待进一步检视，针对深度伪造风险生成机理的研究尚存在明显的“真空”与“留白”。本文基于“技术—资本—权力”的多维视角，结合国内外多个典型案例，采用扎根理论方法考察和揭示深度伪造风险的概念模型及其生成机理，旨在为深度伪造风险的有效治理提供经验证据。

二、研究方法的选择与数据来源

（一）研究方法

目前，关于深度伪造风险的概念模型及其生成机理的研究尚未形成成熟的变量范畴和理论模型体系。为此，本文采用扎根理论对相关案例进行系统性研究。扎根理论由社会学家 Barney Glaser（巴尼·格拉泽）和 Anselm Strauss（安塞尔姆·斯特劳斯）于1967年共同提出，其初衷在于“弥合理论研究与实际经验研究之间的鸿沟”[⑦]。自提出以来，该理论在学术界引起了高度关注，并被广泛应用于管理学、社会学、政治学等多个研究领域。扎根理论的核心方法论强调避免研究者在研究过程中受到主观因素的干扰，防止研究者预先设定假设并仅寻找支持该假设的证据，从而导致研究结果偏离实际。简言之，“扎根理论主张从社会过程及其研究中客观地呈现研究问题，并以此为基础生成理论成果，确保研究结论的真实性与可靠性”[⑧]。由此可见，扎根理论方法在社会科学领域有着广阔的应用前景与独到优势。

本文采用程序性扎根理论对深度伪造风险的概念模型与生成机理进行研究，主要基于以下两点考量：一是深度伪造风险作为一种全球性风险，其复杂性远超单一案例所能涵盖的范围。通过扎根理论，能够整合全球范围内的多样化案例，形成结构化、全景式的分析框架，从而增强研究结论的解释力和普适性。二是扎根理论尤其适用于探索深度伪造风险这类“复杂的新生事物”[⑨]。深度伪造风险并非单一因素作用的结果，而是“技术—资本—权力”等多重因素复杂互动、相互交织的产物。扎根理论方法的引入有助于系统性地剖析和揭示其内在机理，深度全面地揭示其概念模型与生成逻辑，从而深化对深度伪造风险的过程性理解，为相关

---

① Qidong Huang & Jie Zhang, et al., “Initiative defense against facial manipulation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 2, 2021, pp. 1619-1627.

② 龙俊、王天禹：《人工智能深度伪造技术的法律风险防控》，《行政管理改革》2024年第3期。

③ Naffi Nadia & CéLODIE Melodie, et al., “Empowering youth to combat malicious deepfakes and disinformation: An experiential and reflective learning experience informed by personal construct theory,” *Journal of Constructivist Psychology*, Vol. 38, No. 1, 2025, pp. 119-140.

④ 米里亚姆布·伊顿、亚历山大·德斯特里尔、马丁·佩茨等：《法律与经济学下的人工智能责任》，《中国刑事司法》2024年第3期。

⑤ 王璇、宋春龙：《基于深度伪造技术的“二创”视频伦理风险及规制治理研究》，《西南民族大学学报（人文社会科学版）》2024年第5期。

⑥ Shenhao Cao & Xiaohui Liu, et al., “A review of human face forgery and forgery-detection technologies,” *Journal of Image and Graphics*, Vol. 27, No. 4, 2022, pp. 1023-1038.

⑦ 景怀斌：《扎根理论编码的“理论鸿沟”及“类故理”跨越》，《武汉大学学报（哲学社会科学版）》2017年第6期。

⑧ 吴肃然、李名荟：《扎根理论的历史与逻辑》，《社会学研究》2020年第2期。

⑨ 贾哲敏：《扎根理论在公共管理研究中的应用：方法与实践》，《中国行政管理》2015年第3期。

研究提供更为扎实的理论基础。

（二）案例甄选与数据来源

鉴于深度伪造风险具有显著的全球性特征，本文以联合国教科文组织（UNESCO）于2021年发布的《人工智能伦理问题建议书》和国际电信联盟（ITU）于2021年发布的《全球网络安全指数（GCI）2020年版》等国际组织的权威报告为基础，参考其中关于深度伪造风险的威胁评估标准及其分类框架，如“虚假信息、身份盗窃和网络欺诈”① 等具体风险类型，以及“全球性风险、地区性风险、针对部分正常国家的风险、针对脆弱国家的风险”② 等风险层级划分，对深度伪造风险进行系统性分析。

本文选取国际互联网上知名媒体发布的30余篇相关风险报道作为研究样本，其中大部分报道影响力覆盖全球五大洲，浏览人数超过百万，点击量和转发量累计过亿，具有显著的代表性和影响力。秉持典型性和相关性原则对初始样本进行了初步筛选：首先，剔除内容重复或相关性较低的报道；其次，将类型相同的案例进行合并处理，最终保留12篇最具代表性的案例，并运用Nvivo11.0数据分析软件对其加以编号（见表1）。

表1　深度伪造风险的相关案例

序号	地区	名称	来源
1	美国	瓦拉斯支持警察滥用职权视频事件	纽约时报、华盛顿邮报等
2	马来西亚	马来西亚经济事务部长性丑闻视频事件	路透社、海峡时报等
3	印度、巴基斯坦	印巴边境冲突因深度伪造视频升级事件	英国广播公司、半岛电视台等
4	加蓬	加蓬总统视频风波与军事政变事件	华盛顿邮报、英国广播公司
5	美国	叙利亚电子军黑客攻击美联社推特引发金融震荡	路透社、英国广播公司等
6	英国	语音深度伪造诈骗案致英国公司巨额损失	华尔街日报、福布斯等
7	中国	深度伪造版小米董事长“雷军”骂街事件	中央广播电视台、美国有线电视新闻网、英国广播公司等
8	美国	盖尔·加朵遭换脸伪造色情视频事件	美国有线电视新闻网、卫报、泰晤士报等
9	韩国	首尔大学“N号房”深度伪造色情视频案	中央广播电视台、中央日报、美国有线电视新闻网、英国广播公司等
10	美国	美国将深度伪造技术投入于现代战争特种作战	纽约时报、华盛顿邮报、防务新闻等
11	美国	泰勒·斯威夫特遭深度伪造不雅照风波	华盛顿邮报、英国广播公司、美国有线电视新闻网等
12	中国	院士遭换脸进行网络直播带货	中央广播电视台、新华通讯社等

其中，选取前10个案例作为分析基础（序号1—10），还预留2个案例用于饱和度分析（序号11—12）。在此基础上，采用多源数据采集方法，文本分析材料主要来源于经过严格筛选的新闻报道，同时积极整合政府部门网站等官方渠道发布的相关声明与政策文件。在资料收集过程中，综合运用文本研究法、可视化定量法等多种研究方法，并通过三角互证法对数据进行交叉验证，以增强研究结论的科学性、准确性和可靠性。

三、基于扎根理论的资料编码与理论饱和度检验

（一）扎根理论分析的开放式编码

“开放式编码”（Open Coding）作为扎根理论构建的关键起始环节，其核心操作过程包括对原始资料的深度剖析与系统整理，最终实现资料的概念化。具体而言，“研究者需要对转录文本进行逐句精读，通过语句概念化将原始语句提炼为具有代表性的初始概念”③。在此过程中，研究者既可以使用原始文本中的词汇或短语作为初始概念，也可以根据研究需要创造新的概念标签。

本研究严格遵循上述原则，首先对前10个案例进行开放式编码分析，共提取出54个初始概念。随后，

① 联合国教科文组织（UNESCO）：《人工智能伦理问题建议书》，2021年11月25日，https://unesdoc.unesco.org/ark:/48223/pf0000379924；国际电信联盟（ITU）：《全球网络安全指数（GCI）》（2020），2021年6月29日，https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2021-PDF-E.pdf。

② 刘天旭：《脆弱国家研究的兴起：现状、原因及局限》，《国外社会科学》2012年第6期。

③ Anselm Strauss & Juliet Corbin, *Basics of qualitative research: Grounded theory procedures and techniques*, Sage, 1990, pp. 61-74.

基于研究主题对相关报告中的语义场进行系统分析，通过以下步骤对初始概念进行整理：第一，剔除与研究主题无关的概念；第二，合并语义相近或交叉的概念；第三，将相关概念归入同一类别。经过系统梳理与整合，最终形成27个具有明确特征的初始范畴，如表2所示。

表2 开放式编码①

范畴	初始概念	原始资料（范例）
A1 使用成本（a1，a2）	a1 数据成本	“最新的深度伪造应用程序、工具和平台为非法攻击者提供了极为强大的功能，能够以极低的成本快速创建深度伪造视频。”（a1）
	a2 时间成本	“依照行业目前的技术水平，用户只需要找几分钟的零散素材作为学习样本，就能迅速生成出完整的深度伪造视频……”（a2）
……	……	……
C2 素材获取（c3，c4）	c3 目标受众	“浙江大学公共政策研究院的策论指出，虚假的视频或音频可能通过社交媒体迅速传播，误导公众、煽动社会情绪，这里的目标受众就是广大的社交媒体用户，包括各个年龄段、各个社会阶层和不同文化背景的人群”（c3）
	c4 传播特性	“错误信息的传播往往依赖于破坏共识、强调不确定性、削弱领军人物和机构的公信力，以及传播伪科学替代品等方式，而深度伪造信息很容易与这些错误信息相结合，借助其传播模式和渠道，进一步扩大影响范围和危害程度。”（c4）
……	……	……
E3 技术盈利（e5，e6）	e5 数据售卖	“欧盟近年来高科技犯罪案件中，在深度伪造产业链的数据售卖已经成为灰色地带普遍存在的环节，因为制作深度伪造内容需要大量数据，不法分子一般会非法收集、售卖个人数据等给深度伪造的制作者。”（e5）
	e6 技术授权	“深度伪造目前处于AI技术的灰色地带，技术授权缺乏法律依据，盗用、滥用的现象频繁发生。”（e6）
……	……	……
I3 司法证据权威性遭削弱（i5，i6）	i5 伪造证据难辨别	“深度伪造技术可用于伪造视频、音频等证据，且逼真度越来越高，让人难以辨别真伪，给司法等领域带来挑战，干扰正常的司法程序和调查。”（i5）
	i6 真实证据被质疑	“由于深度伪造技术的存在，即使是真实的证据也可能被怀疑是伪造的，影响证据的可信度和法律效力，增加了事实认定的难度。”（i6）

（二）主轴编码

“主轴编码”（Axial Coding）作为扎根理论构建的核心环节，其主要任务是通过建立概念间的关联，将开放式编码阶段形成的离散范畴系统化。具体而言，“研究者需要运用‘范式模型’（paradigm model），以某一核心范畴为轴心，系统分析其与其他范畴之间的逻辑关系，包括因果条件、现象、情境、干预条件、行动策略和结果等要素”②。

在完成开放式编码的基础上进行主轴编码，采用的系统化步骤包括：第一，识别范畴间的因果关系，并分析其内在的互动策略；第二，确定现象发生的情境条件，并识别出干预条件及其影响因素；第三，建立结果与过程间的关联。经过系统分析，最终构建出9个具有明确逻辑关系的主轴编码范式，这些范式不仅深化了对现象本质的理解，也为后续的选择性编码提供了清晰的理论框架，推动研究向更高层次的理论构建迈进（见表3）。

表3 主轴编码③

主范畴	副范畴	关系内涵
技术易用性	使用成本	成本低廉使得深度伪造技术更易普及，恶意使用者可轻松制作虚假内容，扩大其传播范围和频率，进而可能导致虚假信息泛滥，加剧社会危害
	学习门槛	低成本使深度伪造技术极易普及，恶意用户可轻松制作虚假内容，扩大传播，可能导致虚假信息泛滥，加剧社会危害
	素材获取	素材易得提升了深度伪造的效率和质量，伪造者能更快制作逼真虚假内容，加剧危害
……	……	……

① 限于篇幅，本研究仅选取开放式编码的部分情况予以展示。

② Barney Glaser & Anselm Strauss, *The discovery of grounded theory: Strategies for qualitative research*, Aldine Publishing Company, 1967, pp. 45-49.

③ 限于篇幅，本文仅选取主轴编码的部分情况予以展示。

续表3

主范畴	副范畴	关系内涵
资本流向	国际政治领域	深度伪造可能制造虚假外交事件和领导人言论，引发国际争端，破坏关系和舆论，严重冲击国际秩序和国家间信任
	对外军事领域	深度伪造可能制造虚假军事部署和行动视频，误导敌方决策，甚至引发军事误判，严重威胁国家安全
	娱乐文化领域	深度伪造可能损害文化作品的原创性和真实性，导致文化价值混乱，如伪造经典作品、篡改名人形象，影响文化传承，误导观众和消费者，带来文化传播和审美风险
……	……	……
监管滞后	制度保障	完善制度可规范深度伪造技术研发和应用，明确责任，为风险防范提供依据。无制度保障则使深度伪造行为无约束，风险扩大
	法律规制	法律规制是遏制深度伪造风险的关键，通过明确法规界定和处罚犯罪行为，威慑违法者，减少不法行为，为打击提供法律依据和强制力保障
	监管协同机制	监管协同机制整合各方力量，构建全方位监管体系，应对深度伪造风险的跨领域特点，避免漏洞和重复，提升监管效率和风险防控能力

（三）选择性编码

“选择性编码”（Selective Coding）作为扎根理论构建的高级阶段，其主要任务是通过整合与精炼，从主轴编码形成的范畴中识别出核心范畴，并建立系统的理论框架。具体而言，“研究者需要运用‘故事线’（storyline）方法，以核心范畴为中心，系统整合各范畴间的逻辑关系，构建具有解释力的理论模型”①。这一过程重点关注核心范畴与其他范畴的系统整合，包括理论饱和度的检验、范畴间关系的验证以及理论模型的完善。在完成主轴编码的基础上，本研究开展选择性编码：第一，识别最具解释力的核心范畴，建立核心范畴与其他范畴的系统关联；第二，验证范畴间关系的理论饱和度；第三，整合各范畴形成完整的故事线，完善理论模型的解释力。经过系统分析，最终构建出3个具有理论解释力的核心范畴，这些核心范畴不仅形成了完整的理论框架，还实现了对研究现象的深层解释，标志着扎根理论构建的最终完成（见表4）。

表4　选择式编码

核心范畴	主范畴	结果	关系内涵
技术失控	技术易用性	易用性启动失控：低门槛吸引大规模用户，形成技术应用的初始动能	深度伪造技术的失控式发展为资本积累提供了新的变现手段；为权力主体提供了更为高效的控制工具
	技术隐蔽性	隐蔽性掩护失控：黑箱机制阻碍风险识别，延缓社会响应	
	传播扩散性	扩散性放大失控：网络效应与技术依赖使失控后果呈指数级蔓延	
资本逐利	资本规模	外部性转嫁：社会成本未被纳入逐利计算	资本的逐利性驱动深度伪造技术快速发展；为权力主体提供更多政治献金，以换取政策庇护与市场垄断
	资本动机	短期主义：过度追求规模与流向效率，可能牺牲长期稳定收益	
	资本流向	系统性风险：资本集中度过高导致“大而不倒”的脆弱性	
权力寻租	监管滞后	监管滞后提供土壤：规则缺失使寻租行为难以被界定和惩罚	权力寻租推动深度伪造技术向特定方向发展，成为攫取权力与操控舆论的工具；权力主体“投桃报李”，为资本牟利提供政策倾斜与市场保护，形成权力与资本的深度绑定关系
	权力重构	权力重构打开窗口：结构调整中的权责模糊地带成为寻租温床	
	公信力下降	公信力下降削弱制衡：社会监督失效使寻租行为进入“低风险—高收益”区间	

深度伪造风险的形成与演变主要源于技术易用性提升、资本规模扩张以及监管滞后等多重因素的相互作用，其核心驱动机制可归纳为技术失控、资本逐利和权力寻租三个关键维度，技术失控体现在深度伪造技术的低门槛化和快速迭代，资本逐利表现为市场利益驱动下的技术滥用，而权力寻租则反映了监管体系的缺陷与利益博弈。上述三方面因素相互交织，共同构成了深度伪造风险演化的内在逻辑，其关系结构如图1所示。

① Kathy Charmaz, *Constructing grounded theory: A practical guide through qualitative analysis*, Sage Publications, 2006, pp. 96-122.

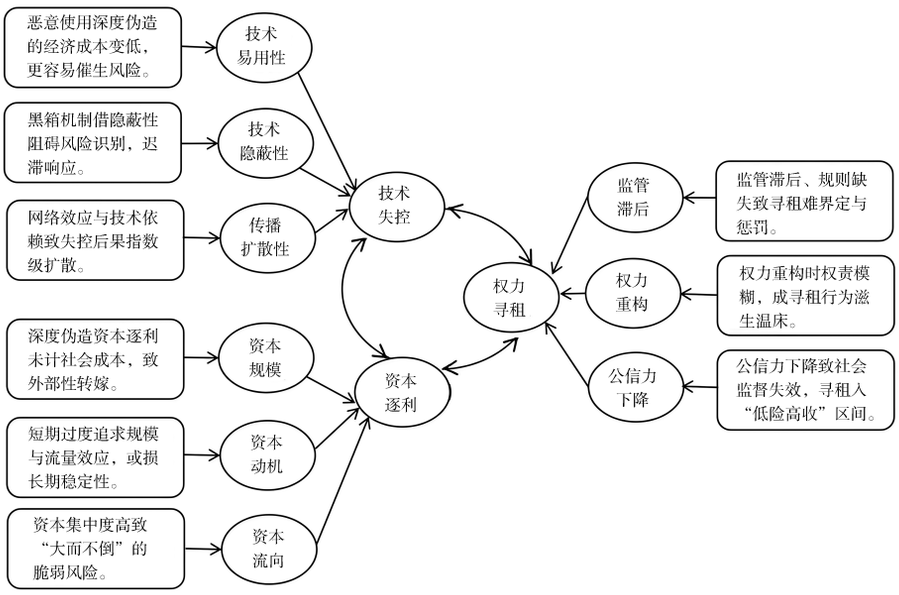

图1 结构关系图

（四）理论饱和度检验

扎根理论的研究过程始于对现象的观察，终于理论饱和的达成。理论饱和是指当新的数据不再能够产生新的理论见解或范畴时，研究即可终止。[①] 这一标准被广泛认为是评估质性研究信度和效度的重要指标。[②] 为验证理论饱和度，本研究采用Nvivo11.0质性数据分析软件对案例11和12进行系统分析。研究严格遵循扎根理论的三级编码程序：首先进行开放式编码以识别初始概念，继而通过主轴式编码建立概念间的联系，最后通过选择性编码整合核心范畴。[③] 经过多轮迭代分析，研究发现既未涌现新的理论范畴，现有范畴内部也未出现新的概念或关系网络。这一发现与Glaser提出的理论饱和标准高度吻合，表明所构建的归因模型具有良好的理论饱和度。[④] 可见，本研究通过三级编码最终形成的结构关系图，符合扎根理论研究方法的科学性要求，具有较强的解释力。

四、深度伪造风险的概念模型与生成机理分析

（一）深度伪造风险的概念模型

基于扎根理论方法，围绕核心范畴和发展故事线，本研究系统构建了深度伪造风险的概念模型。概念模型的构建过程主要包括概念化结构处理和可视化表达两个关键步骤。[⑤] 首先，通过概念化结构处理，将深度伪造风险这一抽象概念转化为结构化的理论框架，明确其核心属性、构成要素及其相互关系。[⑥] 其次，为增强概念模型的可解释性与直观性，研究采用“构图法”（Diagrammatic Representation）对模型进行可视化表达。构图法能够将抽象的理论概念转化为直观的图形结构，从而更清晰地展示各构成要素及其相互作用关系。[⑦] 概言之，深度伪造（Deepfake）风险作为一种数字社会的新型风险，其产生与演变并非单一因素所致，而是源于技术属性、资本动机、资本规模、监管能力、权力结构等多重复杂因素的相互叠加与动态互动。这一风险的形成与演变可以从“技术失控”“资本逐利”“权力寻租”三个核心维度构建其概念模型，并由此揭示其内在逻辑与动态机制（见图2）。

---

①⑤ J. Corbin & A Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory (3rd ed.)*, Sage Publications, 2008, pp. 143-145.

②⑥ K. Charmaz, *Constructing grounded theory: A practical guide through qualitative analysis*, Sage publications, 2006, pp. 96-98.

③ A. Bryant, & K. Charmaz, *The Sage handbook of grounded theory*, Sage publications, 2007, pp. 45-47.

④ B. G. Glaser, *Theoretical sensitivity*, Sociology Press, 1978, pp. 72-74.

⑦ M. B. Miles & A. M. Huberman, et al., *Qualitative data analysis: A methods sourcebook (3rd ed.)*, Sage Publications, 2014, pp. 89-92.

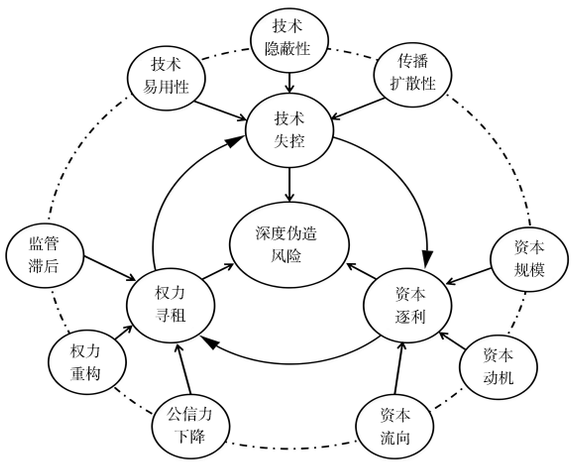

图 2　深度伪造概念模型

其一，技术失控。相较于传统传媒领域的伪造技术，深度伪造技术更加凸显出技术失控的特征，主要体现如下：第一，技术易用性大幅提升。传统伪造技术依赖于专业知识和复杂工具，如照片暗房技术、录音剪辑技术和印刷排版篡改等，这些技术通常由专业传媒机构或技术人员掌握，且因其“以假乱真”的特性而对外保密。相比之下，深度伪造技术借助开源算法和用户友好型软件，使普通人也能轻松生成高度逼真的伪造内容。例如，“a2，依照行业目前的技术水平，用户只需要找几分钟的零散素材作为学习样本，就能迅速生成出完整的深度伪造视频；a4，在社交平台中，深度伪造可以自动化内容创建过程，并允许没有技术专长的普通民众就能轻松操作伪造软件”。第二，技术隐蔽性显著增强。传统伪造技术无论多么精密，总会留下可识别的痕迹，而深度伪造通过计算机神经网络的自我博弈，生成的内容甚至能欺骗其内置的鉴别器，更遑论人类肉眼。例如，“b3，随着人脸生成、人脸属性修改、人脸替换、表情操纵等多种深度伪造工具的应用，使得对伪造视频图像的检测和识别变得越来越困难，仅仅依靠传统视频图像真实性检测和鉴定方法难以支撑多样化的伪造手段。”第三，传播扩散性更加迅速和广泛。深度伪造内容兼具“时间媒介”的持久性和“空间媒介”的广泛性，能够像癌细胞一样迅速扩散。例如，“c3，浙江大学公共政策研究院的策论指出，虚假的视频或音频可能通过社交媒体迅速传播，误导公众、煽动社会情绪，这里的目标受众就是广大的社交媒体用户，包括各个年龄段、各个社会阶层和不同文化背景的人群。”

其二，资本逐利。自诞生之初，深度伪造技术便被众多企业视为数字时代的“迈达斯之触”，被寄予厚望。过去十年间，该技术已广泛应用于影视娱乐、广告营销、时尚美妆等商业领域，并逐渐渗透至灰色地带，成为资本逐利的工具。这一现象的形成主要受以下三方面因素之影响：第一，资本规模的持续扩张引致深度伪造技术的滥用。随着社交平台的迅猛发展，平台竞争日趋白热化。为迎合用户猎奇心理并满足资本扩张需求，许多企业不再局限于合法途径，转而利用深度伪造技术对热点事件进行虚假加工，如明星绯闻、直播带货、国际政治、军事冲突等，并通过精准推送以扩大传播效果。例如，“d2，《卫报》的一项追踪报道指出，部分欧美企业曾利用深度伪造技术伪造网红形象，推荐低劣甚至有害产品，误导消费者购买。”第二，资本对规模与效率的过度追求加剧了技术滥用。在商业竞争压力下，资本方为抢占市场先机，往往不遗余力地寻求竞争优势。深度伪造技术因其处于法律与道德的灰色地带，具有显著的高收益、低风险特征，导致资本方倾向于突破理性决策边界，甚至为追求短期超额利润而牺牲长期稳定收益。例如：“e6，欧盟近年高科技犯罪案件中，深度伪造产业链的数据售卖已成为灰色地带的普遍环节，不法分子通过非法收集、售卖个人数据牟利。”第三，资本流向的过度集中放大了技术滥用风险。随着深度伪造技术在多个产业中的广泛应用，资本逐渐向该领域高度集中，可能形成“大而不倒”的局面。这种资本集中度过高的现象不仅催生了虚幻的经济泡沫，还显著增加了系统脆弱性，一旦泡沫破裂，将引发严重的连锁反应。例如：“e2，Snap 收购 AI Factory，整合其在 GIF 制作等领域的技术优势，虽提升了深度伪造算法的图像质量并拓展了商业应用，但也引发了未来资本是否会过度集中于技术灰色地带的忧虑。”

其三，权力寻租。权力寻租现象正推动深度伪造技术向特定方向发展，使其逐渐异化为攫取权力与操控舆论的工具。这一趋势的形成有其深层逻辑：第一，技术发展与监管滞后之间的结构性矛盾为权力寻租提供了空间。深度伪造技术以其惊人的迭代速度突破了传统监管框架的应对能力，形成了“技术先行、监管追赶”的被动局面。例如：“g3，反网络欺诈公司 Sumsub 统计显示，2022 年至 2023 年全球检测到的深度伪造事件数量增长了 10 倍，而同期相关监管政策的出台数量却远低于此。技术与监管之间的严重失衡为权力寻租创造了制度漏洞，现有法律如美国《通信规范法》第 230 条，常被平台用作逃避责任的工具，进一步加剧了监管失效。”第二，权力重构过程中的制度模糊地带为寻租行为提供了温床。在数字化转型的背景下，传统权力

结构正在经历深刻调整，新的权责边界尚未完全明晰。这种过渡期的制度模糊性，加之深度伪造技术的隐蔽性特征，共同助长权力寻租行为在监管盲区中的肆意滋生与蔓延。例如："g5，欧美国家虽在讨论整合政府资源以协同防治深度伪造风险，但联邦政府与地方政府之间难以达成共识，导致防治措施难以有效落实。"第三，社会公信力的持续下降削弱了制衡机制的有效性，使寻租行为进入"低风险—高收益"的恶性循环。随着虚假信息的泛滥，公众对传统信息源的信任度不断降低，这也为深度伪造内容的传播提供了土壤。例如："h3，《华盛顿邮报》的报道指出，'深度伪造'会加深公众对政府的不信任感，使公众对真实视频产生怀疑，甚至将真实信息误认为虚假信息，对官方澄清也持怀疑态度。"这种信任危机不仅降低了权力寻租的道德成本，也严重影响并削弱了社会监督机制的效用发挥。

（二）深度伪造风险的生成机理

在概念模型的基础上，本研究进一步构建了深度伪造风险的生成机理模型，以揭示其形成机制与演化路径。"生成机理模型旨在解释研究对象（深度伪造风险）从潜在状态到显性状态的关键驱动因素及其作用机制。"① 通过对案例数据的多轮迭代分析，研究发现深度伪造风险的生成机理主要涉及技术失控、资本逐利和权力寻租三个核心维度，三个核心维度的因素通过复杂的相互作用共同催生了风险的生成与扩散，该生成机理模型的构建为深入理解和洞悉深度伪造风险的动态性和复杂性提供了系统性视角，也为深度伪造风险的防控和治理提供了理论依据（见图3）。

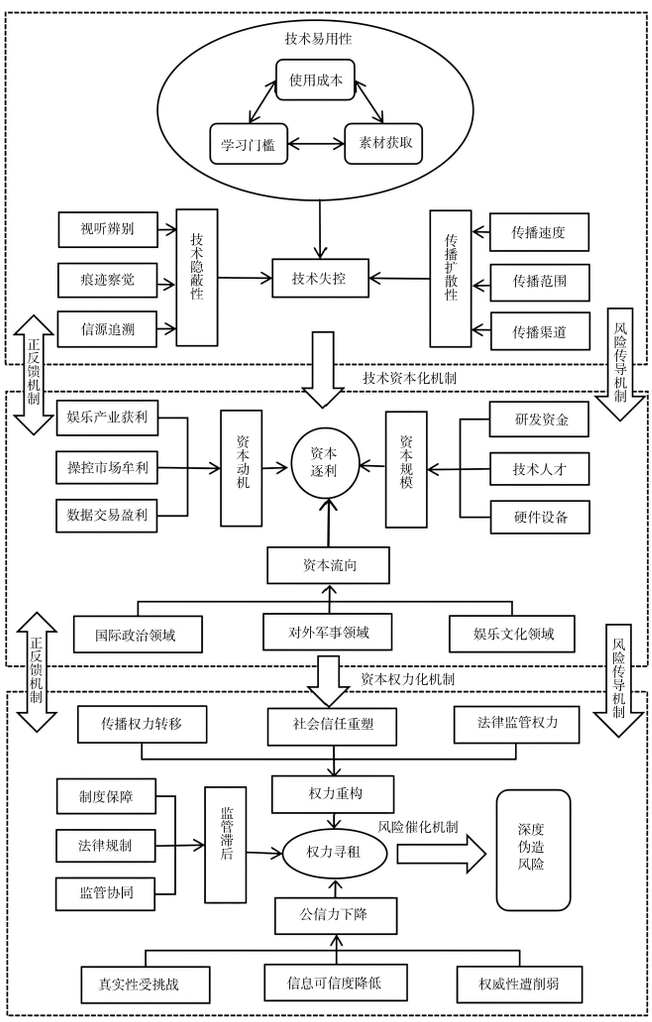

图3 深度伪造风险的生成机理

其一，技术资本化机制。深度伪造技术从"产品"到"商品"的转化过程，本质上体现了技术资本化机制的核心作用。这一过程可划分为两个关键阶段：技术产业化阶段和技术资本化阶段。每个阶段都伴随着特定的经济逻辑和社会条件，共同推动技术从实验室走向市场，最终被资本所主导和控制。在技术产业化阶段，新技术在初始阶段往往面临高研发成本、技术不成熟以及应用场景有限等挑战。然而，"在社会分工体系的框架下，每个生产者的专业化劳动都成为社会总劳动的一部分"②。这种分工协作不仅提升了技术应用的效率，还加速了技术的迭代与优化。"随着新的生产要素（数字技术）组合一旦被领先企业采用，将在产业中扩散，降低成本并促使原先的生产方法标准化。"③ 与此同时，市场需求的潜力被逐步挖掘，应用场景从实验性尝试扩展到商业化落地，从而为商品交换奠定了坚实的基础。这一阶段的核心在于技术从实验室走向市场，

① Jon Elster, *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*, Cambridge University Press, 2015, pp. 107–110.

② Karl Marx, *Capital: A critique of political economy, Volume I*, Penguin Classics, 1992, pp. 461–462.

③ Joseph Alois Schumpeter, *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle*, Harvard University Press, 1934, p. 71.

完成从“概念”到“产品”的跨越。进入技术资本化阶段后，商品交换的属性吸引了资本方的关注。“资本通过投资（深度伪造）技术，进一步扩大生产规模，优化技术性能，并通过市场营销和商业模式创新，将技术推向更广泛的市场。”[①] 在这一过程中，资本的逐利性驱动技术向高利润领域集中，技术的研发方向和应用范围逐渐受到资本的深刻影响。例如，AI“院士”带货表明，深度伪造技术总会被优先应用于娱乐、广告等高回报行业，而其在医疗、教育等社会公益领域的应用则可能被边缘化。最终，技术发展逐渐被资本逻辑所主导，其核心价值从追求社会效益转向追逐经济效益，呈现出明显的工具化倾向。

其二，资本权力化机制。资本权力化机制揭示了资本如何通过深度伪造技术攫取和巩固权力，这一过程体现了资本、技术与权力之间复杂的交织与互动关系。随着“信息技术与金融资本的合谋”日益紧密[②]，这种“合谋”关系不仅改变了劳动与资本的动态平衡，还使得资本逐渐获得了原本属于传统权力的控制力。在资本尚未完全掌握权力之前，权力机构主要通过政策引导和法律监管来规范信息技术的发展方向，确保其创新与应用始终处于合法合规的框架内。与此同时，信息技术也为打破传统的权力分配格局、推动社会关系的变革与进步提供了重要支持，形成了技术与权力之间的良性互动。然而，在资本逐利本性的驱使下，深度伪造技术逐渐偏离其技术中立的初衷，被资本用于操控信息、塑造公众认知甚至干预政治进程。资本方通过掌控信息传播渠道和技术资源，能够构建和形塑有利于自身利益的舆论环境，从而实现对社会的隐性控制。例如，利用深度伪造技术制造虚假政治事件、伪造名人言论或操纵视听内容，以影响选举结果、政策制定或公众情绪，叙利亚电子军黑客攻击事件就是一明证。这种资本与权力的深度融合，使得深度伪造技术从一种技术创新演变为一种新型的权力工具，进一步加剧了社会不平等和权力结构的失衡。资本权力化机制的核心在于，资本通过技术手段实现了对信息生产和传播的垄断，从而获得对社会话语权和决策权的实质性控制。这种控制不仅体现在经济领域，还渗透到政治、文化和社会生活的方方面面，形成了“技术—资本—权力”三位一体的新型统治模式。在这一模式下，深度伪造技术既是资本扩张的工具，也是权力重构的重要媒介，最终导致技术沦为资本权力化的重要推手。

其三，风险催化机制。风险催化机制揭示了深度伪造技术如何在社会系统中催化并放大风险，这一过程不仅体现了技术本身的特性，也反映出技术与社会的复杂互动关系。深度伪造技术的隐蔽性和高度逼真性使得虚假信息难以被识别和验证，从而大幅降低了虚假信息的传播门槛，并加速了其在社会领域中的扩散速度。例如，一段利用“电报”（telegram）等信息平台深度伪造而成的军事政变、自然灾害或政治丑闻的消息可以在短短几分钟内传遍社交媒体，进而引发公众恐慌、社会动荡乃至于国际关系紧张。这种虚假信息的传播所带来的危害具有极强的“涟漪效应”，就如同向“大池塘”中丢下一块巨大的“石块”，能够对社会信任体系造成巨大和恶劣的影响。与此同时，当虚假信息频繁出现且难以辨别时，公众对信息来源的信任度会逐渐下降，甚至对官方澄清和权威机构的声明也持怀疑态度。信任危机不仅削弱了社会共识的基础，还势必导致公众对信息的普遍冷漠或极端化倾向，形成“沉默螺旋”，进一步加剧社会分裂和对立。例如，“瓦拉斯”的选举失败足以表明，伪造的政治人物言论可能引发选民对选举公正性的质疑，从而破坏民主制度的合法性。更为严重的是，深度伪造技术的风险催化作用具有累积效应。“随着技术的不断进步和普及，虚假信息的制作和传播成本将进一步降低，而其社会危害性却可能呈指数级增长。”[③] 这种技术与社会风险的相互作用，使得深度伪造技术不仅成为社会不稳定的催化剂，还可能引发更深层次的系统性风险，如法律体系的失效、社会治理能力的下降以及国际关系的恶化。

其四，正反馈机制。正反馈机制揭示了深度伪造技术滥用如何形成一种自我强化的恶性循环，这一过程不仅体现了技术与社会的动态互动，还反映了技术滥用对系统稳定性的深远影响。“随着深度伪造技术的普及和滥用，社会对虚假内容的辨识能力和警惕性逐渐降低，虚假信息的传播效果因此不断增强。”[④] 这种传播效

---

① Joseph Alois Schumpeter, *Capitalism, socialism, and democracy*, Harper & Row, 1962, pp. 86-87.

② Phil Jones, *Work Without the Worker: Labour in the Age of Platform Capitalism*, Verso Books, 2021, pp. 115-116.

③ Bruce Schneier, *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World (First edition)*, W. W. Norton & Company, 2015, pp. 111-112.

④ Soroush Vosoughi & Deb Roy, et al., “The spread of true and false news online,” *Science*, Vol. 359, No. 6380, 2018, pp. 1146-1151.

果的提升反过来又激励更多资本和权力主体投入资源，进一步开发和应用深度伪造技术，以追求更大的经济利益或政治影响力。例如，伪造的明星带货效果越好，资本方越倾向于投资相关技术以扩大其市场影响力，而政治势力也可能利用该技术操纵舆论或干预选举。这种“技术滥用—传播效果增强—资源投入增加—技术滥用加剧”的正反馈循环，使得深度伪造技术的风险不断累积，最终形成难以控制的局面。在这一过程中，深度伪造技术的滥用不仅加剧了信息环境的复杂性，还对社会信任体系造成了系统性破坏。“随着虚假信息的泛滥，公众对信息来源的信任度逐渐下降，甚至对权威机构的澄清和辟谣也持怀疑态度。”① 这种信任危机进一步削弱了社会共识的基础，导致信息传播的“劣币驱逐良币”现象，即虚假信息因其煽动性和传播效率而更容易被接受，而真实信息则因缺乏吸引力而被边缘化。例如，在重大公共事件中，伪造的视频或音频可能迅速引发社会恐慌，而官方的澄清却难以获得同等关注，从而加剧社会不稳定，印巴边境冲突就是对此最好的例证。最后，正反馈机制还可能造成技术发展的路径依赖。随着资本和权力主体对深度伪造技术的依赖加深，技术研发的方向可能逐渐偏离公共利益，转而服务于少数群体的私利。这种技术发展的“锁定效应”不仅限制了技术的多元化应用，还可能阻碍其他更具社会价值的技术创新。例如，当前深度伪造技术在娱乐和广告领域的过度应用，可能挤占其在医疗、教育等公益领域的研发资源，从而加剧技术的社会风险，削弱技术创新的公共利益属性。

其五，风险传导机制。风险传导机制揭示了深度伪造风险如何在不同领域和社会层面传导与扩散，这一过程不仅体现了技术风险的复杂性，还反映出现代社会系统的高度互联性。深度伪造技术的风险并非囿于技术本身，而是“通过信息传播、社会信任和政治干预等多重途径，向经济、社会乃至国际政治领域传导和扩散，形成一种（‘量子纠缠’式的）‘涟漪效应’”②。在经济领域，深度伪造技术的滥用可能通过虚假信息的传播引发市场波动。例如，“伪造的企业财报、高管言论或突发事件视频可能误导投资者决策，导致股市剧烈震荡或资本外流”③。这种经济风险不仅影响个体企业的声誉和运营，还可能波及整个行业甚至国家经济体系的稳定性。此外，深度伪造技术在金融诈骗中的应用，如伪造身份信息或交易记录，也进一步加剧了金融系统的脆弱性。在政治领域，深度伪造技术的风险传导效应尤为显著。通过伪造政治人物的言论或行为，“深度伪造技术可能加剧国家间的政治抹黑、经济犯罪甚至恐怖主义行动”④。例如，一段伪造的政治领袖演讲视频可能引发国内政治动荡或国际紧张局势，进而影响政策制定和国际合作。这种政治风险的传导不仅威胁民主制度的合法性，还可能引发地缘政治冲突，成为国际关系中令人感到棘手和头疼的不稳定因素。在社会领域，深度伪造技术的滥用对社会信任体系造成了深远破坏。“当虚假信息与真实信息相互掺杂、难以辨认时，公众将会对‘官方’的权威性，乃至于合法性产生怀疑，进而造成‘真相的沙漠化’。”⑤ 这种信任危机不仅削弱了社会共识的基础，还可能导致公众对信息的普遍冷漠或极端化倾向，进一步加剧社会分裂和对立，造成社会鸿沟乃至引发社会冲突。例如，在特种作战中，伪造的专家言论或政府政策将会影响公众对当前作战的看法，甚至最终导致“战争难民潮”，进而影响到一国或多国的公共安全。

五、结论与建议

在深度伪造技术迅猛发展和肆意传播扩散的背景下，深度伪造风险无疑成为公共治理中的一道棘手难题。扎根理论方法的引入和运用，为解读和揭示深度伪造风险的概念模型及其生成机理提供了可行的途径。

第一，深度伪造风险是指基于生成对抗网络所生成的虚假音频、视频与图片等内容，由于其具有高度的

① Sam Gregory, “Deepfakes, misinformation, and disinformation, and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism,” *Journalism*, Vol. 23, No. 3, 2022, pp. 708-729.

② Justin Cochran & Stuart Napshin, “Deepfakes: Awareness, concerns, and platform accountability,” *Cyberpsychology, Behavior, and Social Networking*, Vol. 24, No. 3, 2021, pp. 164-172.

③ Maja Nieweglowska & Cal Stellato, et al., “Deepfakes: Vehicles for Radicalization, Not Persuasion,” *Current Directions in Psychological Science*, Vol. 32, No. 3, 2023, pp. 236-241.

④ 李冲、李霞：《人工智能风险的试探性治理：概念框架与案例解析》，《中国软科学》2024年第4期。

⑤ Mika Westerlund, “The emergence of deepfake technology: A review,” *Technology Innovation Management Review*, Vol. 9, No. 11, 2019, pp. 39-52.

逼真性和易传播性等特性，对隐私权、创作权、财产权等公民权利，以及对社会信任、经济秩序和政治安全等多个领域所构成的潜在威胁。该风险不仅体现在传统信息伪造技术所引发的社会信任危机，还包括技术滥用导致的欺诈、诽谤、舆论操纵等行为，以及由此衍生的法律、伦理和安全挑战。深度伪造风险的核心在于其技术特性与传播机制的结合，使得虚假内容能够快速扩散并产生深远的社会影响。本文突破既有研究的微观视角，以“技术—资本—权力”的整体框架为切入点，为深度伪造风险的研究提供了更为清晰的概念界定和理论依据，也为这一领域开辟了跨学科研究的新视角与新方向。

第二，深度伪造风险的概念模型可从“技术失控”“资本逐利”和“权力寻租”三个维度加以构建：技术失控体现为人工智能生成技术的低门槛与高逼真性，使得虚假内容脱离可控范围；资本逐利驱动技术的商业化应用，在追求利润最大化的过程中加剧技术滥用；权力寻租则表现为技术掌握者通过制造虚假信息操纵舆论、干预政治，甚至制造国际争端，以实现权力扩张或利益攫取。三者之间相互作用，共同构成了深度伪造风险的生成与扩散机制，为理解和认识深度伪造风险提供了系统性框架。

第三，深度伪造风险的生成机理表现为一个多元机制共同作用的动态过程。技术资本化机制作为起点，通过资本投入推动技术的快速发展和应用，为深度伪造技术的普及奠定了基础；资本权力化机制则进一步将技术转化为操纵社会权力的工具，使技术掌握者能够通过制造虚假信息影响公众认知和决策。在此过程中，风险催化机制因技术的低门槛和传播平台的算法操控，加速了虚假信息的扩散，而正反馈机制使得虚假内容在传播中不断自我强化，形成恶性循环。最终，风险传导机制将风险从个人隐私领域一步步扩展到社会、经济和政治领域，继而引发连锁反应，形成多层次、多领域的风险放大效应。这一系列机制的相互作用，共同构成了深度伪造风险生成和扩散的完整链条。

上述研究发现不仅为深度伪造风险的概念模型构建和生成机理解析提供了重要的理论依据与实证支撑，同时也为进一步探索深度伪造风险的治理路径指明了方向。

一要明确立法方向，划定法律底线。深度伪造风险引发的“大洪水”究竟是否会成为重创人类文明的“灭世灾难”，首先取决于法律的“诺亚方舟”是否“牢固”。就目前而言，深度伪造仍然是风险社会中较为特殊且发展较为迅猛的一种风险，世界各国对其到来尚未做出充分的准备。譬如，前述研究案例中的美、英、韩等国在面对突如其来的深度伪造案件时就陷入了无法可依的被动、尴尬局面。为此，“当以前瞻性的视角和严谨的态度明确该领域的立法方向，并制定清晰、可行的法律规范”①。通过划定法律底线，构建一个既鼓励技术创新又维护保障社会稳定秩序的法律框架，以确保深度伪造技术在法治的轨道上健康、有序地发展。

二要强化监管机制，确保执行有力。需要加以指出的是，深度伪造技术在教育、文化、数据安全等多领域均有正面和积极的效用，简单粗暴的“一刀切”难免有“因噎废食”之嫌。但是，如同阿克顿所言“一切有权力的人，都容易滥用权力”②。实践中，深度伪造技术已沦为部分不法者滥用“数维坦”的“权柄”。为确保深度伪造技术的健康发展，维护社会的公正与正义，强化监管机制并确保其有效执行显得尤为关键。这不仅是对技术创新边界的审慎考量，更是对社会秩序与公众利益的坚定守护。通过构建严密的监管体系，严厉打击其非法应用，以引导深度伪造技术朝着积极、正面的方向发展，造福于人类社会。

三要细化制度规范，实现刚柔并济。寻找技术发展与伦理边界之间的黄金分割线是探索深度伪造风险防治之策的风向标。在注重法治和强调监管的向度外，还应积极融入人性化、德治和自治的理念，打造“法治—德治—自治”三位一体的立体化规制体系，实现刚柔并济和协同发力，形成制度合力。同时，针对深度伪造风险治理场域中的多元行动者，要划定清晰的行为边界，以确保多元参与者在享受技术红利的同时承担起相应的社会责任与道德义务。确保规制体系既有刚性约束力，能够有效遏制恶意的伪造行为，以保护社会公众免于虚假信息的侵害；亦不失一定的柔性引导力，以激发技术向善的创新活力，推动深度伪造技术向着合理化、合法化、合规化和人性化的方向发展。

---

① Rebecca Delfino, “Deepfakes on trial: A call to expand the trial judge’s gate keeping role to protect legal proceedings from technological fakery,” *Hastings Law Journal*, Vol. 74, No. 2, 2022, pp. 293–348.

② Lord Acton, *Essays on Freedom and Power*, Beacon Press, 1949, p. 364.

四要加强国际合作，共同应对挑战。“没有人是一座孤岛，可以自全，每个人都是大陆的一片，整体的一部分。”① 在面对深度伪造风险这一世界性和复杂化的治理难题，任何国家都难以独善其身，唯有加强国际合作，方能共同应对挑战和难题。这意味着“首先要在超国家层面建立相应的监管机制，这就需要有条件地制定新条约以明确禁止或控制该技术在国际政治和国家安全领域的发展或使用”②。对此，各国亟需建立跨越国界的壁垒，携手构建深度伪造防治的全球网络，通过共享情报、交流经验、协同立法与技术创新，共同筑起一道跨国界、协同化和立体化的抵御虚假信息的坚固防线。

五要提升公众意识，营造良好环境。“人是目的，而非手段。”③ 无论多么精妙、富有创造力的新技术都应服务于社会大众。事实上，就科技发展史而言，任何一种被大规模运用于社会生产的数字技术，最终都会以不同的形式“飞入寻常百姓家”。人民史观告诉我们，“历史是群众的事业，人类历史发展的本质是人民群众的发展”④。因此，发挥广大人民群众的力量，提升社会公众对于深度伪造技术的认知与防范意识，营造一个基于真实、崇尚理性的社会环境是防治深度伪造技术、维护信息生态安全的关键所在。“只有更多的公众意识到了深度伪造可能造成的危害，进而在整个社会培养防范深度伪造的公众意识，深度伪造对国民安全造成的威胁才能降到最低。”⑤ 如此，方能最大程度地发挥和释放深度伪造技术的积极效应，实现“智能向善，造福人类”的终极目标。

〔本文为国家社会科学基金重点项目“数字赋能促进公共服务高质量供给及其实现路径研究”（21AGL032）的阶段性成果〕

（责任编辑：王胜强）

Conceptual Model and Generation Mechanism of Deepfake Risks from the Perspective of Grounded Theory

ZHAN Guobin, CHEN Yifan

Abstract: In the digital era, deepfake technology pervades cyberspace, undermining the authenticity, transparency, and credibility of public governance through deception and threatening public and national security. Research shows that deepfake risks emerge from a dynamic, multi-mechanism process: technological capitalization kick-starts rapid commercialization; capital's power transformation weaponizes the technology for social manipulation by spreading misinformation; the risk catalysis mechanism is exacerbated by the low technical threshold and the manipulation of platform algorithms; positive feedback loops reinforce false content during diffusion; and finally, the risk transmission mechanism amplifies cross-domain and multi-level harm. Thus, a holistic governance framework integrating legal regulation, policy oversight, institutional standardization, international cooperation, and public participation is essential to ensure the legal and orderly application of deepfake technology in public governance.

Key words: deepfake risks, risk governance, conceptual model, generation mechanism

① John Donne, *Devotions upon Emergent Occasions*, Oxford UP, 1987, p. 86.

② 张远婷：《人工智能时代“深度伪造”滥用行为的法律规制》，《理论月刊》2022年第9期。

③ Immanuel Kant, *Grundlegung zur Metaphysik der Sitten*, Felix Meiner Verlag, 1999, p. 429.

④ Karl Marx & Friedrich Engels, *Die heilige Familie, oder Kritik der kritischen Kritik*, Dietz Verlag, 1959, p. 98.

⑤ 刘国柱：《深度伪造与国家安全：基于总体国家安全观的视角》，《国际安全研究》2022年第3期。