Open Journal of Legal Science 法学, 2025, 13(11), 2475-2481
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2025.1311338

Hans汉斯

生成式人工智能内容的风险及法律规制

李植钧

北方工业大学文法学院，北京

收稿日期：2025年9月29日；录用日期：2025年10月10日；发布日期：2025年11月7日

摘　要

生成式人工智能技术是人工智能领域里的一项突破性技术，其推动着社会生产活动发展的同时，也带来了诸多的新挑战。数据安全、劣质信息泛滥、训练数据来源合法性争议等成为讨论焦点，对这些问题进行法律规制迫在眉睫。本文将对现有法律规制出现的难题进行分析，通过对国内外法律进行比较法的动态研究，学习外国法治的优点；推动技术治理工具的进一步完善，使得生成式人工智能内容的透明度进一步提高。并对技术提供者适用过错推定原则，进一步遏制生成式人工智能内容出现的各种法律问题，有助于推动人工智能领域的健康发展。

关键词

生成式人工智能，虚假有害信息，数据安全，服务提供者，算法黑箱

The Risks and Legal Regulation of Generative Artificial Intelligence Content

Zhijun Li

School of Humanities and Law, North China University of Technology, Beijing

Received: September 29, 2025; accepted: October 10, 2025; published: November 7, 2025

Abstract

Generative artificial intelligence technology is a breakthrough in the field of artificial intelligence, which not only promotes the development of social production activities but also brings many new challenges. Issues such as data security, the proliferation of low-quality information, and the legitimacy of training data sources have become the focus of discussion. Legal regulation of these problems is urgently needed. This paper will analyze the difficulties in existing legal regulations, conduct a comparative study of domestic and foreign laws through a dynamic approach to learn from the advantages of foreign rule of law; promote the further improvement of technical governance tools

文章引用: 李植钧. 生成式人工智能内容的风险及法律规制[J]. 法学, 2025, 13(11): 2475-2481.
DOI: 10.12677/ojls.2025.1311338

to enhance the transparency of generative artificial intelligence content. Moreover, applying the principle of presumed fault to technology providers can further curb various legal problems arising from generative artificial intelligence content, which is conducive to the healthy development of the artificial intelligence field.

Keywords

Generative Artificial Intelligence, False and Harmful Information, Data Security, Service Providers, Algorithmic Black Box

1. 引言

随着生成式人工智能，从早期符号逻辑推理迈向多模态内容创造的跨越式发展，其已经从数据分析工具转变为具有自主创造力的“数字主体”。自美国 OpenAI 公司推出 ChatGPT 这一现象级产品以来，全球科技领域掀起生成式人工智能技术革新浪潮。谷歌、微软等国际科技领军企业持续加大研发投入，推动 AIGC 领域呈现爆发式增长态势。同时国内的生成式人工智能大语言模型不断涌现，技术应用规模不断扩大。

但是技术大规模的创新应用，也带来了前所未有的风险挑战，新增了许多法律风险。生成式人工智能的卓越性能源于其庞大的数据深度训练及复杂的算法机制，但这种技术实现路径中潜藏的数据泄露风险和算法误用可能性已成为引发社会各领域高度警觉的重要议题。具体而言，模型通过超大规模数据样本的持续学习虽然提升了输出精准度，但训练过程中可能涉及敏感信息滥用、数据污染及算法偏见植入等问题，这些技术特性与治理盲区的叠加效应，正不断加剧技术合规性漏洞和伦理争议。本文将聚焦于生成式人工智能在应用阶段所面临的治理挑战以及法律规制风险，通过梳理其风险特征与监管困境进而提出分层递进的监管策略为构建完善的治理体系和促进行业规模提供前瞻性的解决方案。

2. AIGC 概述

2.1. 生成式人工智能的技术背景

生成式人工智能是人工智能领域的重要分支，其核心在于通过算法和模型从海量数据中学习规律，并自主生成具有逻辑性和创造性的新内容，包括文本、图像、音频、视频及代码等。生成式人工智能目前包含的生成式对抗网络和生成式预训练转化器都是由马尔可夫链和隐马尔可夫模型奠定的生成模型理论发展而来。生成对抗网络的提出实现了高质量的数据生成，开启现代生成式 AI 的新阶段，2018 年大语言模型和多模态技术的突破，推动生成式 AI 进入通用化时代。本文将以 ChatGPT 为例对生成式人工智能的基本工作原理作简要剖析，ChatGPT 的生成机制分为四个阶段[1]。第一阶段是数据收集与预处理。主要是完成原始语料的获取与标准化处理，从互联网抓取海量的文本数据，通过分词等技术进行清洗和结构化。第二阶段是预训练。利用自注意力机制对预处理后的数据进行无监督学习，通过上下文捕捉、语言规律建模、参数优化等方法生成初始内容。第三阶段是微调与优化。使用人工标注的高质量对话数据，进行监督微调，让模型学习符合人类偏好的回答模式，再通过奖励模型对内容进行评分。第四阶段是输出控制。内置内容审核机制，对涉及暴力、歧视等有害内容进行实时拦截。

2.2. 生成式人工智能内容的传播

生成式人工智能正以颠覆性力量渗透至千行百业，通过内容自动化生成与智能化创新，驱动产业效率与用户体验的全面升级，其正在多领域应用中重塑产业生态。在新闻传媒领域，呈现出自动化与深度化并行。生成式人工智能在新闻领域实现了速度与深度的平衡。美联社采用人工智能系统，每年自动生成超 5000 篇企业财报报道，效率较人工提升 20 倍[2]。在深度报道方面，封面新闻的“小封智作”平台，整合大数据与多模态生成技术，可基于热点事件自动生成可视化分析报告。AI 虚拟主播更是能实现 24 小时多语种新闻播报。在娱乐创作领域，人机协同新范式正不断巩固。生成式人工智能正重塑内容创作逻辑，网易伏羲的 AI 编辑工具为游戏《逆水寒》生成 30 万条支线剧情，使得玩家互动时长增加 60%。音乐创作上，谷歌的 MusicLM 可根据各类提示生成高质量音频，Suno 平台用户日均创作歌曲超 1 万首[3]。

3. AIGC 的问题现状

3.1. 数据安全风险

在训练过程中使用的数据可能存在准确性不足或系统性偏差，其完整性和合规性难以充分保障，可能使模型输出内容带有误导性或有害性[4]。随着人工智能技术在各行业的深度渗透，数据泄露风险与合规压力显著加剧，一旦关键数据外泄，将对企业运营、行业生态造成直接经济损失及品牌信誉损害。值得注意的是，即便是零散的敏感信息也可能被类似 ChatGPT 的模型通过跨源数据关联分析，重构出涉及国家安全、公共安全及个人权益的情报[5]。对于部署于海外服务器的模型(如 GPT 系列)，输入敏感数据可能触发跨境数据流动风险，进而威胁国家数据安全。研究显示，仅需篡改训练数据集的 0.001%即可显著降低模型可靠性，这种低成本、高隐蔽性的攻击方式进一步放大了数据安全治理的复杂性。

3.2. 劣质信息泛滥

生成式人工智能凭借其强大的算力可批量生成虚假文本、编造不实信息，导致事实性错误泛滥。不法分子则利用深度合成技术伪造音视频内容，实施诽谤、散布谣言、窃取隐私甚至冒名诈骗等违法行为，严重破坏网络生态与社会稳定。此类滥用行为不仅加剧了信息污染风险，还通过私域传播渠道降低违法成本，例如 AI 换脸技术仅需数张照片即可生成高仿真动态视频，被用于伪装熟人诈骗、虚假广告推广等场景[6]。随着生成式 AI 技术加速普及，内容滥用的实施与扩散成本显著降低，但监管与执法难度呈指数级增长。一方面，AI 生成内容逼真度已超越人类辨识能力，传统审核机制难以有效识别伪造痕迹；另一方面，跨国界、跨平台传播特征要求监管体系具备更强的技术响应能力。例如，实时视频会议中 AI 换脸诈骗案例显示，伪造内容已突破图片易造假、视频可信度高的传统认知。这种技术迭代速度与治理能力之间的错位，倒逼治理体系向动态化、智能化升级，亟需构建分级分类监管规则与跨模态检测技术能力。

3.3. 训练数据来源合法性争议

在数据处理过程中，合规要求要求数据处理者采取必要措施保障数据安全、网络安全及个人信息安全，确保用户数据合法使用并得到有效保护。生成式人工智能在我国法律框架下的数据合规风险主要集中于大模型训练阶段的数据来源合法性问题，具体表现为以下三方面风险。

其一，若训练数据涉及公民个人信息，根据《个人信息保护法》第 13 条，处理前需取得信息主体同意。然而，在海量数据训练场景下，要求开发者逐一对用户履行知情同意程序存在现实困难，由此产生的合规风险亟待解决。其二，训练数据若源于受著作权法保护的作品，受限于我国合理使用制度的法定性，文本挖掘等行为难以被认定为合理使用。若未经著作权人授权，可能构成侵权行为。其三，通过爬

虫技术获取网络数据存在双重合规风险[7]：一方面，可能因技术手段干扰目标网站正常运营，违反《网络安全法》第 27 条关于禁止非法获取个人信息的规定；另一方面，若数据来源于声明禁止爬取的网站，可能被认定为侵犯企业数据财产权益，违反《反不正当竞争法(修订草案征求意见稿)》第 18 条关于保护经营者智力成果的规定。

4. AIGC 法律规制的隐忧

4.1. 法律规范的滞后性

生成式人工智能的爆发式发展，标志着人工智能技术从“工具性应用”向“创造性生成”的范式转变。然而，其引发的法律挑战已远超现有法律体系的应对能力，呈现显著的法律规范滞后性特征。首先是技术特性与法律概念的错位。生成式 AI 的“黑箱”特性与自主学习能力，导致传统法律中“行为主体”“过错责任”等核心概念难以适用[8]。例如，AI 创作内容版权归属争议中，现有法律对“创作主体”界定模糊，而《著作权法》未明确 AI 生成内容是否构成“作品”。其次是风险预防与事后追责的失衡。现有法律多依赖“损害结果发生后的追责”，但生成式 AI 的风险具有隐蔽性与扩散性，如虚假信息传播、数据偏见强化。例如，2023 年美国大选中的 AI 伪造候选人视频事件，暴露了法律在事前风险防控上的空白。最后是跨国监管冲突与治理碎片化。各国立法步调不一：欧盟《人工智能法案》强调高风险场景的透明度要求，中国《生成式人工智能服务管理暂行办法》侧重合规审查，而美国采取行业自律模式。这种差异导致跨国平台面临“监管套利”困境，如数据本地化要求与全球服务模式的冲突。

4.2. 内容审核的复杂性

生成式内容的监管面临多重治理困境。由于算法内部运作机制的不透明性及决策逻辑难以溯源的技术特性，系统决策的生成路径往往超出人类的理解范畴。正如 ChatGPT 在官方说明中强调，其大规模语言模型产出的文本规模呈现指数级增长态势，导致传统人工筛查机制存在效率瓶颈。OpenAI 的专项研究显示，尽管 GPT-4 通过强化训练在内容可信度层面取得突破性进展，但其核心技术短板尚未根除。这种技术迭代带来的表面可靠性提升，反而可能加剧用户使用过程中的认知依赖，致使风险防范意识弱化[9]。此外，对抗性样本与模型迭代也带来了压力。AI 生成工具可通过对抗性训练规避审核算法检测，如调整文本语法结构或图像像素分布。内容审核系统需持续更新模型以应对新型生成技术，但如高性能 GPU、分布式存储系统等硬件升级和算法研发成本高昂，中小平台难以负担。时间成本与用户体验的平衡也加剧了审核的复杂性。复杂审核流程延长内容发布时间，例如传统审核需几分钟，而 AI 内容审核可能耗时数十分钟，导致用户流失。平台需在审核精度与效率间权衡，进一步加剧运营压力。

4.3. 责任主体的复合性

传统网络违法行为通常呈现二元格局(网络平台与受侵害方)或多元参与结构(网络平台、使用者与受侵害方) [10]。前者指网络服务商直接实施违法行为的场景，后者指用户借助网络服务实施侵害的情形。在生成式人工智能应用场景中，权责主体的复杂性体现为两个维度：首先涉及人工智能实体的法律人格争议性，即其是否具备独立法律地位并承担相应责任。学界对此存在对立观点，既有主张赋予有限主体资格的理论探讨，也有坚持工具论立场的反对声音。其次表现为多元潜在责任方的识别难题。相较于传统网络侵权中明确的直接实施主体，生成式人工智能的内容产出机制具有多重交互特性——其内容生成依赖算法架构、算力支撑、数据资源与人机协同。由于算法黑箱效应与可解释性缺陷，各参与要素对最终侵权结果的作用权重难以精确量化。从事实层面观察，算法开发者、数据持有者、终端使用者均可能对违法内容的产生形成关联性影响，呈现出责任溯源困境。

5. AIGC 法律规制的完善进路

5.1. 比较法的动态研究

美国采取的是产业导向的“合理使用”原则。美国以《版权法》的“合理使用”为核心，强调技术创新的优先性。通过“转换性使用”标准，允许对作品的非表达性部分进行算法训练。司法实践中倾向于支持技术行业，例如引入“非表达性合理使用”概念，允许 AI 模型对受版权保护内容的非核心部分进行复制，以平衡市场替代效应与创新需求。2023 年拜登政府要求对 AIGC 添加水印，旨在区分人类与 AI 生成内容，但尚未形成统一的版权保护规则。欧盟模式为权利导向的严格监管[11]。欧盟《数字单一市场版权指令》(DSM 指令)要求商业用途的 AI 训练需获得权利人授权，并设立“文本与数据挖掘例外”，允许科研用途的免费使用，但权利人可通过“选择退出”保留权利。《人工智能法案》强化透明度义务，要求生成式 AI 提供者披露训练数据来源，确保版权人可追溯主张权利。强调对生成内容的标识义务，防止公众混淆 AI 生成物与人类作品，例如要求显著标注深度合成技术生成的内容。而中国则是混合模式下的探索。《生成式人工智能服务管理暂行办法》(2023 年)要求对 AI 生成内容进行标识，但现有规定(如《互联网信息服务深度合成管理规定》)仅覆盖部分场景，存在技术适用性不足的问题。司法实践中，北京互联网法院审理的“AI 生成图片侵权案”尝试界定 AI 生成物的版权属性，但目前尚未明确其是否构成著作权法保护的“作品”。

5.2. 技术治理工具的嵌入

中国在生成式人工智能治理方面通过立法与政策创新构建了多层次监管框架，旨在平衡技术创新与安全风险。2023 年 7 月发布的《生成式人工智能服务管理暂行办法》是该领域的核心法规，其针对生成式 AI 技术通用性、价值观属性及可引导性等特点，从内容合规、算法透明、数据安全等维度提出明确要求，例如禁止生成危害国家安全或社会稳定的信息，并要求在算法设计、数据标注等环节防止歧视性输出。该法规与既有《网络安全法》《数据安全法》《个人信息保护法》形成衔接，同时引入分类分级监管、安全评估、算法备案等机制，既强调技术向善的伦理导向，又通过“包容审慎”原则鼓励行业创新。为应对生成式 AI 训练数据滥用、知识产权侵犯及隐私泄露等风险，政策要求数据来源合法化、标注流程规范化，并推动公共数据开放与高质量训练资源建设。技术治理层面，通过“预训练–微调–对齐”全流程监管，强化模型输出的无害性，并探索基础模型与应用服务的分层治理路径。此外，中国积极参与国际规则制定，推动技术标准与伦理规范的全球协同，同时在司法实践中逐步完善数据权属认定与侵权责任划分，为通用人工智能的法治化发展奠定基础。

5.3. 生成式人工智能提供者适用过错推定原则

生成式人工智能技术产品本质上是生成式人工智能技术，或者说是技术下的算法模型，如果按照技术逻辑去划定承担生成式人工智能产品的责任主体，那么生成式人工智能技术本身应当成为履行义务和承担责任的主体。《生成式人工智能服务管理暂行办法》中规定生成式人工智能产品提供者(提供者既包括个人，也包括组织)承担该产品生成内容生产者的责任，是将生成式人工智能产品提供者视为生成内容的生产者。

生成式人工智能服务提供者的侵权责任认定在当前法律框架下呈现出复杂性与特殊性，需结合技术特性与法律规则综合考量。根据民法典及相关司法实践，生成式人工智能服务提供者的责任认定不适用传统网络服务的“避风港原则”，因其技术特性使其直接参与内容生成过程，而非仅提供存储或传输服务。例如，杭州互联网法院在审理用户生成奥特曼侵权图片案件时指出，生成式 AI 服务兼具技术服务与

内容供给双重属性，其责任认定需动态评估服务性质、技术能力、侵权信息明显程度及后果等因素，合理界定注意义务范围[12]。在过错认定上，法院通常要求服务提供者建立有效的事前防范与事后应对机制，如设置关键词过滤、完善投诉举报渠道、显著标识生成内容等，若未尽到合理注意义务则需承担赔偿责任。

从法律适用角度，生成式 AI 服务提供者可能涉及承揽合同责任与一般过错责任。民法典将低风险生成式 AI 服务类比为承揽合同关系，服务提供者作为“承揽人”需对生成内容负责，尤其在用户输入与 AI 生成高度关联的场景下。而判别式 AI 服务则因其技术特性难以直接适用产品责任或特殊归责原则，需回归一般过错责任框架，结合个案判断服务提供者是否存在过失。例如，广州互联网法院在“大模型侵权首案”中强调，服务提供者未履行风险提示义务或未能及时删除侵权模型，即构成过错。此外，行政监管与侵权责任的衔接也影响责任认定，违反《生成式人工智能服务管理暂行办法》等公法义务可能成为侵权过错的判断依据。

《生成式人工智能服务管理暂行办法》中虽然提到了“服务提供者”的概念，但是并未对其进行定义，在法律责任中也只进行了原则性要求，[1]未对具体行为的法律责任进行细化和规范。笔者认为，《办法》应明晰规制对象边界，厘清与《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律、行政法规的关系和承担责任的边界，以实现在权利和责任规则适用上有序衔接和有机联动。

6. 结语

在人工智能时代，生成式人工智能的迅猛发展正不断重塑社会形态，其适用领域与现实挑战受到了学术界和实务界的广泛关注与讨论。特别是在实践中数据安全、劣质信息泛滥、训练数据来源合法性争议等问题更是成为了研究的重点。对此必须清楚认识法律规范的滞后性、内容审核的复杂性、责任主体的复杂性。本文对生成式人工智能的技术背景和传播领域进行基本介绍，探讨了法律规制的方法，但期待着对于更深层次理论的探索。法治的科学化、合理化是一个漫长的过程，在未来的司法实践中，不断推进法律逻辑与技术逻辑的深度融合，构建治理新范式，定能朝着法治化国家的目标迈进。

参考文献

[1] (2022) OpenAI. https://openai.com/blog/chatgpt

[2] 澎湃新闻. 请注意！这条新闻是腾讯机器人写的[EB/OL]. https://m.thepaper.cn/newsDetail_forward_1373723, 2025-09-13.

[3] 智东西. “音乐 ChatGPT”融资 1.25 亿美元, 秒做爆款神曲, 超 1000 万人已使用[EB/OL]. https://mp.weixin.qq.com/s/Wur_qU-HpGoTAOzMavCA1w, 2025-09-12.

[4] 支振锋. 生成式人工智能大模型的信息内容治理[J]. 政法论坛, 2023, 41(4): 34-48.

[5] 皮勇, 张凌寒, 张吉豫. ChatGPT 带来的风险挑战及法律应对[N]. 人民检察, 2023(07).

[6] 朱嘉珺. 生成式人工智能虚假有害信息规制的挑战与应对——以 ChatGPT 的应用为引[J]. 比较法研究, 2023(5): 34-54.

[7] 斜晓东. 论生成式人工智能的数据安全风险及回应型治理[J]. 东方法学, 2023(5): 106-116.

[8] 毕文轩. 生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J]. 比较法研究, 2023(3): 155-172.

[9] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学, 2023(4): 29-43.

[1]《办法》第二十条：“提供者违反本办法规定的，由网信部门和有关主管部门按照《网络安全法》《数据安全法》《个人信息保护法》等法律、行政法规的规定予以处罚。法律、行政法规没有规定的，由网信部门和有关主管部门依据职责给予警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停或者终止其利用生成式人工智能提供服务，并处一万元以上十万元以下罚款。构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任。”

[10] 徐伟. 论生成式人工智能服务提供者的法律地位及其责任——以 ChatGPT 为例[J]. 法律科学(西北政法大学学报), 2023, 41(4): 69-80.

[11] 张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学, 2023, 45(3): 108-123.

[12] 孙祁. 规范生成式人工智能产品提供者的法律问题研究[J]. 政治与法律, 2023(7): 162-176.