

生成式人工智能数据风险的差异化监管路径研究

——基于“公共数据两级市场”模型

张 珊, 隋立夫, 宋雪溢, 刘 涛, 向宁钰

辽宁科技大学经济与法律学院, 辽宁 鞍山

收稿日期: 2026年8月26日; 录用日期: 2026年9月9日; 发布日期: 2026年10月12日

摘 要

生成式人工智能技术以其强大的内容生成能力在各类数据利用场景中快速渗透, 并深刻改变着数据要素的配置逻辑与治理机制。在释放巨大生产力的同时, 其数据源的不透明性、生成过程的不可控性以及内容输出的不可靠性, 也引发了多层次的数据安全风险。立足生成式人工智能的发展现状, 系统梳理其数据风险的多元缘起, 聚焦数据利用层的数据源片面性、模型运算层的算法隐患、内容产出层的信息失真, 深入分析了信息不公、虚假信息生成与市场信息污染等风险类型。溯源分析表明, 风险的生成与扩散贯穿数据采集、模型训练与内容输出全流程, 公共数据的公共属性与安全敏感性决定了政府行政部门必须承担主导监管责任。在此基础上, 进一步提出构建“公共数据两级市场 + 数据差异化监管”模型, 从一级市场源头管控与二级市场全流程合规等维度出发, 形成适应性强、操作性高、前瞻性足的行政监管体系, 以期实现生成式人工智能数据治理由被动响应向主动治理、由单一规制向差异化协同的范式转型, 从而确保数据利用合法合规与技术创新的动态平衡, 推动人工智能行业可持续发展。

关键词

生成式人工智能, 数据风险, 差异化监管, 公共数据两级市场, 行政监管

A Study on Differentiated Regulatory Pathways for Data Risks in Generative Artificial Intelligence

—Grounded in the “Two-Tier Public Data Market” Model

Shan Zhang, Lifu Sui, Xueyi Song, Tao Liu, Ningyu Xiang

School of Economics and Law, University of Science and Technology Liaoning, Anshan Liaoning

Received: August 26, 2026; accepted: September 9, 2026; published: October 12, 2026

Abstract

Generative artificial intelligence (GenAI) technology, with its powerful content-generation capabilities, has rapidly penetrated various data-utilization scenarios and profoundly reshaped the allocation logic and governance mechanisms of data factors. While unleashing substantial productivity, its opaque data sources, uncontrollable generation processes, and unreliable output content have also triggered multi-layered data security risks. Based on the current development of generative artificial intelligence, this paper systematically examines the diverse origins of its data risks, focusing on the one-sidedness of data sources at the data-utilization layer, the algorithmic hazards at the model-computation layer, and information distortion at the content-output layer, and provides an in-depth analysis of risk types such as information inequality, the generation of false information, and market information pollution. A source-tracing analysis reveals that the generation and diffusion of risks run through the entire process of data collection, model training, and content output, and that the public attributes and security sensitivity of public data determine that government administrative departments must assume the primary regulatory responsibility. On this basis, the paper further proposes the construction of a “two-tier public data market + differentiated data regulation” model, which, from the dimensions of source control in the primary market and full-process compliance in the secondary market, forms an administrative regulatory system that is highly adaptive, highly operable, and forward-looking. This is intended to achieve a paradigm shift in GenAI data governance from passive response to proactive governance and from single-form regulation to differentiated and coordinated regulation, thereby ensuring a dynamic balance between lawful and compliant data utilization and technological innovation, and promoting the sustainable development of the artificial intelligence industry.

Keywords

Generative Artificial Intelligence, Data Risks, Differentiated Regulation, Two-Tier Public Data Market, Administrative Regulation

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 背景与问题提出

2022 年, DalleE2 和 StableDiffusion 等具备图像生成能力的生成式人工智能模型的问世, 标志着人工智能迈入了一个崭新的时代。同年 11 月底, OpenAI 公司推出 ChatGPT, 仅在线上两个月内, 注册用户便突破了一亿, 成为生成式人工智能发展史上的一个重要里程碑应用[1]。2025 年 2 月, DeepSeek 因其低成本和强大功能而受到广泛关注, 掀起了继 ChatGPT 问世后又一次全球性的生成式人工智能热潮。生成式人工智能之所以如此引人注目, 是因为其利用从互联网获取的海量训练数据开展深度学习和强化学习, 以数据、算力、算法共同协作做出内容。对于生成式人工智能来说, 训练数据的重要性不言而喻。数量庞大且质量上乘的训练数据是构建高效、准确模型的关键。

公共数据是指各级机关单位、企事业单位依法履职或提供公共服务过程中产生的数据。《全国数据资源调查报告(2025 年)》显示, 2025 年, 全国年度数据生产总量达 52.26 泽字节, 同比增长 27.28%, 占全球约 27.44%, 数据赋能人工智能迈入规模化应用新阶段, 用于人工智能训练和推理的数据总量同比增

长 42.86%¹，推理数据量首超训练数据量。我国数据要素市场化价值化进程显著提速，数据资源供给体系、数据流通体系和数据开发利用体系协同演进，正由数据资源规模扩张向数据要素价值释放跃升。但公共数据一旦被不当利用或监管失序，不仅可能放大上述风险，还可能危及公共利益与国家安全。因此，如何在释放生成式人工智能技术红利的同时，有效防范和化解数据风险，成为亟待回应的治理命题。

各国生成式人工智能服务的监管治理体系在重创新还是重监管的基本理念，以及在政府监督、自监督或是第三方监督的模式选择中都呈现出显著差异^[2]。在此背景下，学界现有对生成式人工智能的监管模式主要分为四类，包括敏捷治理、韧性治理、回应型治理和险基治理。其中，任敏提出的回应性监管体系，主张合理分配监管权，通过“三方主义”引入第三方团体来破解监管困境^[3]；高翔构建了以敏捷治理为基础的复合型监管体系，形成由网信办牵头的横向多部门联合、纵向分层级授权形态，纳入多元协同机制^[4]。上述研究虽明确了分级分类监管路径，但存在共性缺陷：重事后风险管控、轻数据源头治理，且监管责任分散模糊，难以充分应对数据风险。

要破解上述治理困境，有必要回溯至更为基础的理论层面加以审视。从数据治理的理论谱系来看，公共物品理论揭示了数据的非竞争性与非排他性，公共数据尤具公共资源属性，这构成了政府介入的基本理据。然而，数据经加工治理后可以成为排他性商业资产，这种“公私二元属性”决定了单一的市场化或完全政府管控均非最优解。外部性理论表明，数据规模化利用产生正外部性，而泄露、歧视与虚假信息等问题则构成负外部性，需要借助制度安排将外部性内在化。信息不对称理论进一步指出，数据利用链条中存在多重信息鸿沟，使市场自律与事后追责难以独立奏效。委托代理理论则为公共数据授权运营中政府与运营主体之间的权责配置与激励相容问题提供了分析框架。上述理论从正当性论证与制度构造层面为政府监管提供了依据，但均止于“为何监管”的正当性论证，未能针对生成式人工智能的特殊治理场景形成具体的制度方案。在数据要素市场化的制度探索层面，张素华、吴雯琛采用数据价值链理论将公共数据授权运营全流程解构为数据采集、数据汇集、数据治理、数据利用四个子环节。“一级市场 + 二级市场”结构被视为公共数据市场化配置的理想框架。然而，既有研究多着眼于一般性的数据交易与授权运营，尚未将生成式人工智能训练数据的特殊性纳入分析视野，亦未能将差异化监管工具内嵌于市场结构之中。

基于上述理论梳理与既有研究的不足，本文构建“公共数据两级市场 + 数据差异化监管”模型，其核心逻辑在于：依据数据利用的风险等级与技术链条的不同环节，将监管资源差异化配置于两级市场之中。该模型的理论贡献体现为三个方面：其一，将公共物品理论所要求的政府责任具体化为一级市场的源头管控职能，同时通过二级市场的市场化运营兼容数据的商品性特征，实现公共物品理论的制度转化；其二，在风险等级分类基础上增加“技术环节”维度，构建“风险等级 × 技术环节”的二维分析框架，拓展差异化监管的分析深度；其三，将监管功能前置性地嵌入市场结构设计，使市场本身成为风险治理的制度载体，实现从“外部规制”向“内生治理”的范式转换。为此，研究者结合现有模型的优劣，搭建“公共数据两级市场 + 数据差异化监管”模型，通过事前源头分级与事后动态监管双轮驱动，精准补齐数据风险治理短板，为生成式人工智能监管提供全链条行政法解决方案。

2. 生成式人工智能数据风险的类型化识别

根据《生成式人工智能服务管理暂行办法》²第二十二条规定：生成式人工智能技术，是指具有文本、图片、音频、视频等内容生成能力的模型及相关技术。不同于传统人工智能仅对输入数据进行处理和分析，生成式人工智能可以学习并模拟事物的内在规律，根据用户的输入资料生成具有逻辑性和连贯性的

¹<https://dzswgf.mofcom.gov.cn/news/43/2026/6/1780379327067.html>

²https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

新内容，其依托大型自然语言处理模型，能够实现非结构化多模态信息的深度解析与结构化输出[5]。生成式人工智能遵循“数据采集-模型运算-内容产出”的技术链条逐层展开，每一环节均内嵌特定的风险形态。依据其运行机理，可将数据风险类型化为三个维度，表现为获取阶段的合法性风险、使用阶段的泄露风险和生成阶段的内容失范风险。

2.1. 数据获取的合法性风险

训练数据获取同时面临个人信息保护与著作权保护的双重约束，而单纯依赖概括授权或事后补救均难以形成稳定预期。

第一，个人信息处理合法性问题。《中华人民共和国个人信息保护法》³(后文简称《个人信息保护法》)第6条规定：“处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式。收集个人信息，应当限于实现处理目的的最小范围，不得过度收集个人信息。”生成式人工智能的训练依赖海量数据，这些数据来源繁杂，既包括自行采集的内容，也涵盖从公开网络抓取的信息，且同一数据可能在平台、标注机构与模型开发者之间多次流转。然而，在数据规模庞大、技术链条复杂的现实下，逐一征得用户同意、清晰说明训练目的并设置有效的退出机制，均面临重重困难。为提升模型性能，平台往往过度收集数据，轻易突破“最小必要”原则；同时，将公开数据视作可无限利用的资源，也消解了信息主体对后续处理的合理预期。更值得警惕的是，在平台与用户地位不对等的背景下，所谓形式上的同意，往往难以代表真实、自愿且具体的授权。

第二，训练数据利用的著作权风险。算法对作品进行分析、抽取、分词和重组，虽然未必表现为传统意义上的完整复制，却可能影响复制权、改编权和信息网络传播权等。权利人难以识别、授权成本高以及利用范围不确定等风险，均使传统许可机制难以适应模型训练。现行《中华人民共和国著作权法》⁴(后文简称《著作权法》)关于合理使用的列举规则和“三步检验法”，亦难以直接回应数据挖掘等新型利用。如果坚持逐项许可，可能使训练成本和交易成本过高；如果完全否定权利人的控制权，又会损害其创作激励和作品市场。

2.2. 数据使用的泄露风险

数据泄露风险贯穿模型训练、运行和输出全过程。

第一，模型训练容易造成数据泄露风险。用户向模型输入的个人信息、隐私和商业秘密，可能在交互、存储或后续训练中被再次使用。经过参数化处理后，数据虽不以原始形式存在，却可能以关联信息、记忆片段或特定输出的形式重新出现。模型提取攻击表明，攻击者可以从模型输出中还原部分训练数据、个人可识别信息或代码片段。其根源在于，用户通常只能通过格式化协议接受平台的数据处理安排，难以实质控制数据用途，而模型的持续学习和参数化存储又使删除、脱敏和回溯难度增加。

第二，模型运行环节参与主体的多元化加剧了数据泄露的风险。开发者、标注服务商及其他技术提供者之间的数据共享，扩大了接触主体和泄露面。韩国三星公司引入 ChatGPT 后发生的多起信息泄露事件以及用户对话数据和付款服务支付信息丢失情况⁵，都说明此类风险并非抽象假设。跨境处理还可能涉及重要数据、国家机关数据和国防建设数据，泄露后果也因而超出一一般隐私侵害范围。若开发者未能区分训练集、验证集和交互数据，也未对第三方服务商设置访问边界，泄露风险还会沿供应链继续扩散。

第三，模型输出环节也可能潜藏数据泄露风险。即便输入经脱敏处理、运行过程严格隔离，模型仍可能在推理时拼合训练记忆中的碎片信息，还原出个人隐私或商业秘密。攻击者可借助多轮问答拼接或

³http://www.npc.gov.cn/c2/c30834/202108/t20210820_313088.html

⁴https://www.gov.cn/guoqing/2021-10/29/content_5647633.htm

⁵<https://web.csia.net.cn/newsinfo/5722774.html>

成员推理等手法，间接获取非公开内容；精心设计的提示词亦可能绕过安全对齐机制，迫使模型输出敏感结果。此类风险隐蔽性强，开发者和使用者常因侧重输入与运行安全而忽视输出侧隐患。因此，部署后须持续监测生成内容，建立动态过滤与脱敏机制，方可有效降低泄漏可能。

2.3. 数据生成内容的失范风险

生成内容的安全性取决于训练数据的质量。训练数据的内容失范主要表现为以下三个方面。

第一，数据选择、清洗和标注包含价值判断，来源区域、文化背景及标注人员差异会使偏见进入数据集，并在模型输出过程中被放大。标注规则不清、工具不可靠或管理不足，还会使错误分类持续进入训练过程。

第二，网络数据真假混杂，商业目标和算法架构可能使虚假信息得到保留。即使训练数据真实，人类反馈强化学习也可能使模型生成形式合理而实质错误的内容。偏差一旦进入模型，便可能形成难以定位的系统性失真。

第三，失范数据会使生成内容出现歧视、仇恨和深度伪造等问题，增加其责任认定和损害评估难度。训练数据还可能包含涉黄涉赌等不良内容，使模型在特定提示下生成违法或有害内容。

由此，内容风险并非单纯的输出审查问题，其根源仍在于训练数据的选择、处理和反馈机制；若只删除最终结果，则难以阻断偏见和虚假信息在模型内部的累积。特别是当模型生成内容再次成为网络数据或训练材料时，系统可能形成由错误信息反向塑造模型的循环。

综上，生成式人工智能的数据风险并非孤立存在，而是沿着技术链条逐级传导、叠加放大。这种风险的类型化识别，为后续针对性治理措施的提出奠定了认知基础。

3. “公共数据两级市场 + 数据差异化监管”模型构建

张素华、吴雯琛在提出数据价值链理论，将公共数据授权运营全流程解构为数据采集、数据汇集、数据治理、数据利用四个子环节[6]。在此基础上，本文提出的“公共数据两级市场 + 数据差异化监管”模型，正是对上述理论框架的承接与拓展。该模型的核心逻辑在于：依据数据利用的风险等级与技术链条的不同环节，将监管资源配置于两级市场之中，实现差异化、精准化的治理目标。

3.1. 行政监管的必要性

生成式人工智能数据风险的公共属性，决定了行政监管不可替代的主导地位。公共数据的来源主体是各级机关单位与公共企事业单位，其本质属性决定了数据安全不仅关涉个体权益，更与国家利益、公共利益深度绑定。单纯依赖市场自律或技术自治，难以有效应对数据源不透明、算法黑箱与内容失范等系统性风险。数据运营主体的逐利性与公共数据的公益性之间存在天然张力，这一矛盾构成了政府全面监管的内在理据。当前，制度供给滞后与治理需求复杂化的矛盾日益突出，授权机制混杂、监管机制不完善等问题亟须通过行政权力的制度化介入加以破解。因此，由行政机关主导构建覆盖数据全生命周期的监管框架，既是防范风险扩散的制度底线，也是释放公共数据价值的必要保障。

3.2. 一级市场监管路径：源头管控

一级市场以公共机构为主导，具有公益性，负责数据采集、汇集与治理环节，形成公共数据资源池。这一阶段的风险主要集中于数据获取的合法性与数据质量的可靠性，监管目标在于从源头阻断风险的导入。

其一，建立训练数据分类分级登记制度。并非所有公共数据均具有同等的安全敏感性，应根据数据类型、涉密等级与潜在风险程度实施差异化的准入管理。对于涉及国家安全、核心公共利益的高风险数

据，应设定更为严格的审批与脱敏要求；对于一般性公共数据，则可在标准化流程下简化准入程序。训练数据的来源、用途与处理方式应当向监管部门备案，确保数据来源可追溯、用途可核查。

其二，明确数据获取的合规边界。针对个人信息处理合法性问题，应在《个人信息保护法》确立的最小必要原则基础上，结合生成式人工智能规模化训练的特殊需求，探索建立“场景化授权 + 动态同意”机制，而非一概否定大规模数据利用的合法性空间。针对著作权风险，应明确区分“以复制为目的的使用”与“以训练为目的的非表达性使用”，为数据挖掘等新型利用方式提供合理的制度接口。

其三，建立数据质量审核与偏见筛查机制。一级市场应承担数据“净化”的功能，对进入资源池的公共数据进行系统性的事实核查、偏见检测与质量评估。数据选择、清洗与标注环节应制定统一的技术标准与操作规范，避免因标注人员差异或规则不清导致偏见进入数据集。对已识别的偏见与错误数据，应建立标记、隔离与修正的闭环处理流程。

3.3. 二级市场监管路径：全流程合规

二级市场以市场主体为主导，具有营利性，聚焦数据利用环节，开发面向社会提供的公共数据产品。这一阶段的风险涵盖模型训练中的泄漏风险、算法运算中的偏见与黑箱风险，以及内容产出中的虚假信息

其，实行授权运营的准入与退出机制。参与二级市场的运营主体须通过公开遴选方式确定，授权期限与范围应与数据风险等级相匹配。运营机构应具备相应的数据安全能力与技术保障条件，并接受定期评估。对于严重违反数据安全规定或持续输出有害内容的主体，应建立强制退出与从业限制机制。

其二，构建全流程安全审计与监测体系。二级市场的监管不应停留于事后的结果审查，而应嵌入模型训练、运行与输出的全过程。应建立覆盖数据输入、模型运算与内容产出的实时监测与审计溯源机制。对训练数据的每一次调用、模型的每一次迭代、内容的每一次输出均应留有可追溯的日志记录，确保风险事件发生后能够准确定位责任环节。同时，应强化授权运营域的行为审计、在线备案与异常处置能力。

其三，建立内容产出的事前审核与事后追责机制。针对生成内容的虚假信息、深度伪造与知识产权风险，二级市场的运营主体应承担内容安全的主体责任。在技术层面，应部署自动化内容审核与风险预警系统；在制度层面，应建立用户投诉、内容标记与快速处置的联动机制。对于已产生损害后果的违法或有害内容，应能够通过审计溯源路径准确追溯至具体的训练数据批次与模型版本，为责任认定提供证据支撑。

其四，完善收益分配与公益保障的平衡机制。二级市场的营利性特征可能诱发运营主体在数据利用中过度追求商业回报而忽视公共安全。应在授权运营协议中明确公益优先原则，设定数据产品开发的底线安全标准，并通过收益分配机制的设计引导运营主体在合规框架内开展创新。

3.4. 两级市场的衔接与差异化监管的逻辑展开

差异化监管的核心逻辑在于：风险等级决定监管强度，技术环节决定监管方式。在一级市场，监管重在“入口把关”——通过分类分级登记、合规审查与质量审核，从源头阻断高风险数据的流入；在二级市场，监管重在“过程管控”——通过准入退出、安全审计与内容审核，实现对数据利用全流程的动态覆盖。两个市场的监管功能各有侧重又相互配合，共同构成覆盖生成式人工智能数据风险全链条的制度屏障。

一级市场与二级市场并非彼此割裂，而是通过公共数据授权运营平台实现有效衔接。一级市场完成数据的“净化”与合规化处理后，通过平台向二级市场供给标准化的数据资源；二级市场在此基础上开展模型训练与产品开发，其运营过程与产出结果又通过平台的监测系统反馈至一级市场，形成闭环监管链条。

4. 结论

生成式人工智能的迅速发展在释放数据要素价值、推动技术创新和产业升级的同时，也带来了显著的数据安全风险。尤其是立足于我国国家治理体系与结构下，数据安全不仅关乎网络空间的清朗秩序，更直接涉及公共利益保护、国家安全保障与社会治理效能。数据风险不再是传统意义上的“技术附属问题”，而是关涉数据要素市场化配置、数字治理体系现代化以及人工智能健康发展的核心议题，亟须纳入人工智能整体治理战略的关键层面。本文正是基于这一认识，围绕生成式人工智能的数据利用过程，系统探讨了数据风险的差异化监管路径。

研究从数据采集、模型运算与内容产出三个层级系统梳理了生成式人工智能数据风险的生成逻辑，厘清了数据源、算法机制与输出行为等多因子耦合互动下的风险链条，揭示了信息不公、虚假信息生成与市场信息污染等多维风险类型的特征、机制与影响。同时，在风险生成与扩散机制上，不仅涉及技术内生局限、数据源片面性和开发者价值取向偏差，也与数据采集规范缺失、算法审查缺位、监管机构协同不足及法律约束滞后等制度性因素相关。特别是公共数据的公共属性与安全敏感性，决定了政府行政部门必须承担主导监管责任，而不能单纯依赖市场自律或技术自治。

由此，应对生成式人工智能数据风险，本文结合国内外数据治理经验与发展趋势，立足我国实际情况提出“公共数据两级市场 + 数据差异化监管”模型，从一级市场源头管控与二级市场全流程合规等维度构建系统性应对策略，强调形成包容审慎、动态适配、多元协同的行政监管机制，推动数据风险治理从“事后补救”向“事前预防”与“过程嵌入”转型，实现数据利用全流程合规。两级市场通过公共数据授权运营平台衔接，形成闭环监管链条，使政府监管功能在两级市场中实现差异化配置和协同联动。

可以预见，生成式人工智能仍将以指数级速度演进，其数据利用的边界与风险形态尚难以准确预测，而治理能力的提升却远未同步。未来，应持续推进跨学科交叉研究，加强数据伦理、法律制度与技术发展的同步耦合，构建中国特色、全球视野的生成式人工智能数据治理体系，在保障数据安全与公共利益的同时，充分释放技术创新的积极效能，真正实现技术进步与社会福祉的动态平衡。

参考文献

- [1] 于水, 范德志. 新一代人工智能 ChatGPT 的价值挑战及其包容性治理[J]. 海南大学学报(人文社会科学版), 2023, 41(5): 82-90.
- [2] 陈少威, 杨涛, 贾开. 比较政策研究视野下全球人工智能治理模式的差异、共识与改革启示[J]. 中国行政管理, 2024, 40(12): 15-24.
- [3] 任梅, 代栋栋. 生成式人工智能的回应性监管架构及其实现路径[J]. 中国行政管理, 2025, 41(2): 39-50.
- [4] 高翔, 闫钰琪. 我国生成式人工智能服务的复合型监管治理框架及其优化——基于政策文本和深度访谈的混合研究[J]. 电子政务, 2025(5): 1-15.
- [5] 亓蕾. 人工智能训练数据的法律风险与制度供给[J]. 人民司法, 2025(11): 11-19.
- [6] 张素华, 吴雯琛. 公共数据授权运营的模式检视与完善路径[J]. 电子政务, 2026(3): 68-79.