浙江理工大学学报,第56卷,第1期,2026年2月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(s).2026.01.013

大模型合规监管适应性治理的法治因应

周春光
(西华大学知识产权学院,成都 610039)

摘　要:大模型快速发展在推动技术创新的同时,会逐步加深对海量数据的依赖,也因此潜藏着隐私泄露、算法偏见等诸多风险。技术迭代速度远超法律适应能力,信息不对称与利益不均衡问题日益凸显,进而导致大模型合规监管暴露出法律规制体系缺失、监管主体职权不明、应对挑战能力不足、国际规制分歧与合作困境等问题。为应对大模型合规监管滞后问题,需从完善大模型合规监管的法律法规、明确合规监管的主体与职责、统筹技术能力与合规标准、加强大模型生态的国际合作与交流等维度入手,实现大模型合规监管适应性治理的范式转型。

关键词:大模型;合规监管;治理;信息不对称;法治因应
中图分类号:DF411.91　　**文献标志码**:A　　**文章编号**:1673-3851 (2026) 02-0104-10

Legal response to the adaptive governance of large model compliance regulation

ZHOU Chunguang
(School of Intellectual Property, Xihua University, Chengdu 610039, China)

Abstract: The rapid development of large models, while promoting technological innovation, will gradually deepen the dependence on massive data, which implies many risks such as privacy leakage and algorithmic bias. With the speed of technology iteration far exceeding the adaptability of law, information asymmetry phenomenon and unbalanced protection of interests are becoming increasingly prominent, which in turn leads to large model compliance regulation exposing the lack of legal regulatory system, unclear authority of the regulatory body, insufficient capacity to address compliance challenges, and challenges arising from divergent international regulations and cooperative difficulties. Based on the overall appearance of the lagging compliance regulation of large models, it is necessary to achieve the paradigm shift of adaptive governance of large model compliance regulation by improving the laws and regulations of large model compliance regulation, clarifying the subjects and responsibilities of large model compliance regulation, integrating the technological capabilities and compliance standards of large models, and strengthening the international cooperation and exchange of large model ecosystems in multiple dimensions.

Key words: large model; compliance regulation; governance; information asymmetry; rule of law response

收稿日期:2025-07-09　　网络出版日期:2025-12-02
基金项目:四川省哲学社会科学规划课题(SC24FZ028);四川省哲社重点研究基地社会发展与社会风险控制研究中心项目(SR23A06);四川省社会治理创新研究中心项目(SHZLZD2405);四川警察执法研究中心项目(JCZFQN202403);西部乡村振兴研究中心项目(WRR202313)
作者简介:周春光(1989—　),男,陕西汉中人,研究员,博士,主要从事数字法学、规制法学方面的研究。

在国家战略引领与政策驱动的背景下,人工智能领域正经历以大规模预训练模型为核心的技术革新。截至 2024 年 11 月,我国已有 35 家中央企业和国有企业成功部署 66 个大模型[1]。这些模型技术不仅在国内呈现蓬勃发展态势,更在全球范围内加速迭代与部署,标志着该领域已进入国家级力量主导的系统性突破阶段。毋庸讳言,大模型合规监管不仅是保障人工智能安全、可靠、可信发展的核心制度保障,也是推动技术红利向社会普惠价值有效转化的关键治理路径。

然而,既有的法律体系主要建立在传统技术风险应对语境下,其规则设计存在一定的滞后性,难以有效适配大模型技术动态演进、跨界融合所带来的复杂治理需求。此类制度供给与技术现实之间的张力,导致诸多衍生风险持续游离于监管视野的灰色地带,不仅加剧技术伦理失范、市场竞争失序与用户权益受损的系统性隐患,也凸显出既有治理框架在应对新兴人工智能治理挑战时的局限性。因此,完善大模型合规监管的法治路径,推动其从静态合规向适应性治理范式转型,已成为当前亟待深入回应与开展制度创新的重要命题。

目前,我国学界围绕大模型合规监管的研究主要聚焦于监管理念、现实挑战与应对策略三方面。在监管理念方面,有学者主张秉持包容审慎、分层分类、深入细致与快速响应的原则,推动谨慎而灵活的制度尝试[2],亦有观点强调,发展与安全并重应作为我国科技法治建设的核心导向与基本原则[3]。就合规监管所面临的挑战而言,有学者指出隐私泄露、算法偏见等伦理风险在大模型应用中日渐显现[4],也有学者关注到大模型可控性不足所引发的技术治理悖论[5],凸显监管与技术发展间的适配难题。在应对措施层面,有学者建议通过加快标准体系建设以强化技术支撑[6],依托大算力、大模型等技术突破,构建与之相适应的法律规制框架[7],实现“以技术赋能监管”的治理目标。总体而言,尽管我国学界已在大模型监管的基本理念、挑战与对策等方面展开探讨,但现有研究仍多集中于宏观层面的理论探讨,专门从法律规制角度切入的研究尚显不足,导致相关理论建构与制度回应难以充分适配大模型的技术特质与治理需求。鉴于此,本文试图从法治维度回应大模型的合规监管需求,从理论厘清与规范检视、现实困境剖析以及法治因应路径三个层面展开论述,以期为构建契合大模型技术特质的法律治理体系提供学理支撑。

一、大模型合规监管的理论厘清与规范检视

对大模型合规监管予以理论厘清,有助于把握其理论基础与逻辑起点;对大模型合规监管的规范进行检视,则能够释明法律规制的整体外观。通过“实践认知—理论深化—规范评估”这一完整逻辑链条,可为廓清大模型合规监管的现实困境提供理论与规范维度的准备。从核心概念界定来看,大模型是拥有庞大参数与复杂计算架构的深度学习模型[8],其设计初衷在于增强模型的表征能力与预测效能,以适配复杂任务场景及高维数据挑战。大模型合规监管借助规范技术开发与应用行为,防范其潜在风险,从而保障大模型安全、可信、负责任地发展。

(一)大模型合规监管的理论解构

大模型合规监管从现象透析迈向范式重构需要理论分析工具的支撑。对大模型合规监管予以理论解构,能够厘清其内在逻辑与规制要素,从而为制度的理性设计与体系化安排奠定坚实的理论基础。

1. 信息不对称理论

信息不对称理论旨在分析经济决策过程中,各主体之间信息分布的不均衡现象及其所产生的行为影响[9]。在大模型合规监管框架下,决策质量与执行意愿的首要瓶颈在于数据可及性与完整性不足,而治理体系的结构性缺陷多源于系统性资料局限与信息分布失衡。此种信息生态之下,监管主体难以进行全景式感知与精准判断,反而可能刺激被监管方通过扭曲信息来规避监管约束,最终导致规制失灵,具体可从三方面展开分析:

首先,信息鸿沟显著增加了捕捉适宜监管窗口的难度。大模型发展初期,技术和应用尚在探索之中,此时过早监管可能抑制创新活力、限制技术潜力,对监管环境塑造产生负面影响。然而,立法与规制行为的天然滞后性同样暗藏风险。在人工智能高速发展的背景下,若未能适时构建有效的监管框架,大模型的应用可能陷入结构性困境,致使多维风险持续累积并在临界点爆发,极端情况下甚至引发技术路径系统性偏离,进而触发宏观经济波动与社会治理危机。

其次,信息结构的高度碎片化及内在失真风险,使监管主体陷入认知局限。大模型所依赖的算法黑箱及其跨场景应用的异质性,导致任何单一信息源均难以提供全景式、低噪声的数据基础,这种分散且可信度存疑的信息输入,一定程度上削弱了监管评

估的准确性与可靠性,进一步增加了监管决策的难度。

最后,受规制主体的策略性信息隐匿行为,会进一步加剧信息不对称。为降低合规成本,大模型产业链中的核心企业可能主动遮蔽敏感参数、操纵披露内容或伪造校准指标,在监管视野中构筑失真的运营图像,导致行政机构无法准确识别真实风险敞口与技术缺陷,难以实现有效监管。值得注意的是,信息不对称理论为大模型合规监管提供了破局思路。例如,信息披露水平和市场关注程度的提高可以显著消减企业与市场之间的信息偏差。在大模型监管场域中,通过提升算法透明度、完善强制性披露机制与构建多主体协同的监督网络,同样能有效抑制因信息不对称导致的决策偏差和不公平现象,为大模型合规监管的制度设计指明方向。

2.利益相关者理论

利益相关者理论主张,企业在追求股东利益最大化的同时,也应关注其他利益相关者的权益与长期价值创造[10]。该理论突破了传统的股东至上范式,强调企业通过平衡多方利益实现可持续发展。在构建法治化治理体系时,需警惕其异化为阻碍智能技术发展的制度性壁垒。

近年来,我国智能产业依托技术革命迅速崛起,实现指数级跃升,既巩固内需版图,亦在全球价值链中占据重要地位。然而,产业高速扩张并未消弭深层结构性隐忧。原生理论供给匮乏、关键核心技术对外依存度较高,叠加域外技术围堵与市场挤压,这些因素共同制约着产业稳健发展。因此,监管范式应强化赋能与定向扶持功能,借助精细化的制度工具弥合产业链的薄弱环节。当前治理结构侧重于规则制定与执行,对激活企业创新动力的考量相对不足。在生成式智能技术快速迭代的情境下,部分企业因市场竞争激烈,倾向于在研发与商业化进程中控制风险治理成本。若监管框架未与企业内控目标形成耦合,且缺乏正向合规激励,技术主体可能选择策略性规避监管,优先追逐利润与市场份额,进而促使高风险产品与服务流入市场。而利益相关者理论为大模型合规治理提供了规范基础,其倡导多元协同、共享治理与责任分担机制,有助于构建更加公平、透明和可持续的监管体系。

(二)大模型合规监管的规范检视

截至2025年4月1日,通过系统整理相关规范文本,可以归纳出现行法律规制的整体框架,从而明晰大模型合规监管规范文本的体系轮廓。

1.国家层面关涉大模型合规监管的规范文本检视

从立法的层级性来看,我国大模型技术规范文本的制定主体涵盖全国人民代表大会常务委员会(以下简称"人大常委会")、国家互联网信息办公室(以下简称"网信办")、工业和信息化部(以下简称"工信部")、公安部等多个立法与行政监管主体,形成了多部门协同治理的格局。其中,《中华人民共和国网络安全法》(以下简称《网络安全法》)、《中华人民共和国数据安全法》(以下简称《数据安全法》)、《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)构成了该领域的基础性法律,作为上位法为大模型合规监管提供了核心法律依据,并与下位规范形成紧密衔接的制度逻辑。在具体监管规则层面,与大模型合规监管直接相关的规范性文件核心包括《互联网信息服务算法推荐管理规定》(以下简称《算法推荐管理规定》)、《互联网信息服务深度合成管理规定》(以下简称《深度合成管理规定》)和《生成式人工智能服务管理暂行办法》(以下简称《服务管理暂行办法》)等规范性文件(见表1)。

通过整理国家层面与大模型合规监管相关的规范文本,不仅可以从外观识别不同规范的层级安排,还能从立法的针对性、完备性、精准性与系统性深化认知。

从立法的针对性来看,我国大模型监管框架主要集中于网络安全、数据安全与个人信息保护三大领域,核心法律依据包括《数据安全法》和《网络安全法》,立法延伸至与大型模型紧密相关的算法服务和深度合成服务等领域。《算法推荐管理规定》和《深度合成管理规定》等相关法规相继制定并实施,既进一步明确服务提供者的主体责任,又强化用户知情权保障,体现出一定的针对性。然而,当前技术迭代加速、应用场景不断拓展,整体规范体系仍存在滞后性与涵摄性不足的问题。

从立法的完备性来看,现行规范体系中倡导性规定居多,而具有强制约束力的禁止性条款则相对不足;且规范缺乏可操作的具体落实机制,也未为权利受损的主体建立清晰的救济路径。

从立法的精准性来看,《网络安全法》《数据安全法》《个人信息保护法》作为基础性法律,其规定侧重于原则性的宏观导引,规范密度较低,难以直接应对大模型发展中的具体问题。而效力层级较低的规范性文件虽意图细化规则,但其权威性与约束力有限,实践中面临可操作性不足的困境。

表 1　国家层面与大模型合规监管相关的规范文本

文件名称	施行日期	规范位阶	制定主体	相关内容
《中华人民共和国网络安全法》	2017.06.01	法律	全国人大常委会	网络运营者及 AI 系统需保障网络免受干扰、破坏与非法访问,防止数据外泄、被盗或篡改。存储与传输数据时,必须采取技术与必要措施,确保网络安全稳定。
《中华人民共和国数据安全法》	2021.09.01	法律	全国人大常委会	数据处理者需完善全流程安全管理制度,采用技术措施保障数据安全。AI 领域的数据收集、使用、共享等均受此限。
《中华人民共和国个人信息保护法》	2021.11.01	法律	全国人大常委会	处理个人信息须合法、正当、必要、诚信,保障个人知情权、决定权。AI 应用于个人信息时,必须确保个人信息的有效保护。
《互联网信息服务算法推荐管理规定》	2022.03.01	部门规章	国家互联网信息办公室、工业和信息化部、公安部、国家市场监督管理总局	算法推荐服务提供者应落实安全主体责任,建立健全管理制度和技术保障,定期审核内容。保障用户知情选择权,明示算法原理,提供关闭或删除标签功能。针对特殊群体制定保护条款,分级管理,留存网络日志并配合监管。
《互联网信息服务深度合成管理规定》	2023.01.10	部门规章	国家互联网信息办公室、工业和信息化部、公安部	深度合成服务提供者须依法许可或备案,建立管理制度和技术保障,明确服务标识,完善辟谣和投诉举报机制。服务者和技术支持者不得制作、传播违法信息,不得利用服务从事危害国家安全、侵犯权益等犯罪活动。
《生成式人工智能服务管理暂行办法》	2023.08.15	部门规章	国家互联网信息办公室、国家发展和改革委员会等七部门	生成式 AI 服务商需获许可或备案,遵守法律伦理,尊重版权。应制定数据标注、算法、安全评估等制度,明确服务标识,保障用户知情权。严禁生成违法信息,设立投诉渠道,快速响应问题。
《人工智能伦理安全风险防范指引》	2021.01.05	部门规范性文件	全国信息安全标准化技术委员会	提供人工智能伦理安全风险防范措施,明确研究开发者、设计制造者、部署应用者的责任,强调可解释性、可控性,保障人工智能安全可控,促进健康发展。
《新一代人工智能伦理规范》	2021.09.25	部门规范性文件	国家新一代人工智能治理专业委员会	倡导人工智能遵循福祉增进、公平公正、隐私安全保护、生命尊重及可控性原则,规范研发应用伦理,引导技术向人类价值观靠拢,防范算法歧视,保护弱势群体,规避伦理风险。
《生成式人工智能服务内容标识方法》	2023.08.25	部门工作文件	全国信息安全标准化技术委员会	规定了生成式人工智能服务提供者如何为所生成的内容进行标记或识别的具体方法。
《科技伦理审查办法(试行)》	2023.12.01	部门规范性文件	科学技术部、教育部、工业和信息化部等十部门	各单位在进行涉及科技伦理的研究时,须建立完善的伦理审查机制,对相关科研活动实施伦理审查,AI 科研活动涉及伦理问题,均需按规定审查。

从立法的系统性来看,我国大模型合规监管的相关规范已散见于法律、部门规章及规范性文件等多个方面。当前的人工智能立法呈现出显著的应用场景导向,治理主体针对不同产业情境中的特定风险,逐步形成了以专项规制为核心的应对路径。例如,《服务管理暂行办法》侧重回应版权、伦理与安全挑战;而《深度合成管理规定》则为深度合成技术的应用设定了合规框架。此类聚焦型立法在实践中已显现其治理效能。然而,由于缺乏一部统领性的基础法律,现有监管体系存在覆盖盲区,难以系统性地适应人工智能技术的快速迭代,其整体性与前瞻性仍有待加强。

当前,人工智能立法筹备已纳入国家议程。根据《新一代人工智能发展规划》战略部署,中国计划 2025 年初步建立人工智能法律法规、伦理规范和政策体系,展望 2030 年建成更加完善的人工智能法律法规、伦理规范和政策体系,这充分反映出国家对人工智能领域的高度重视。

2. 地方层面关涉大模型合规监管的规范文本检视

截至当前,我国大模型领域的立法实践呈现出立法分散化与地方情境适应性的显著特征。从国家立法层面观察,针对大模型这一特定领域,尚未出台专门性的法律或行政法规,相关制度供给主要由地方依据区域发展实际需求开展先行先试,以探索适配性治理路径。地方层面与大模型合规监管相关的规范文本见表 2。从地方层面看,已出台的大模型相关立法文件效力位阶普遍较低;内容上以促进性、倡导性条款为主,尚未形成体系化的规制框架,整体未能满足该领域对规范密度的现实需求。仅一部设区的地方性法规《成都市数据条例》,主要针对数据处理提出要求,难以有效衔接并系统性地应用于大模型合规监管的复杂进程。

表2 地方层面与大模型合规监管相关的规范文本

文件名称	施行日期	规范位阶	制定主体	相关内容
《北京市促进通用人工智能创新发展的若干措施》	2023.05.23	地方工作文件	北京市人民政府	进行大模型创新算法与关键技术的研发工作,加大大模型训练数据的采集力度并研发相应的治理工具。建立大模型评测的开放服务平台,并构建起大模型所需的基础软硬件支撑体系
《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》	2023.08.04	地方工作文件	成都市经济和信息化局等	促进核心技术的创新突破,建立完善的开发工具体系,塑造大模型技术的产业生态,助力企业主体不断壮大发展
《上海市推动人工智能大模型创新发展若干措施(2023—2025年)》	2023.10.20	地方工作文件	上海市经济和信息化委员会等	开展大模型创新支持项目,设立大模型测试评价中心,推行大模型算力提升计划,打造智能芯片软硬件协同发展的生态环境,实现语料数据资源的共同建设和共享利用,推动大模型创新应用
《广东省人民政府关于加快建设通用人工智能产业创新引领地的实施意见》	2023.11.03	地方规范性文件	广东省人民政府	加大大模型关键技术的研发攻坚力度,强化测评保障技术的研发工作,不断扶持软硬件产品的创新发展,构建算力算法的交易服务平台
《2024年北京市全面优化营商环境工作要点》	2024.04.17	地方工作文件	北京市人民政府	加快构建人工智能全栈创新体系,推动大模型在医疗、政务等领域深度应用,实施包容审慎与分级分类监管模式。同时,建立卓越工程师培育平台,鼓励人才在实战项目中攻坚克难
《成都市数据条例》	2024.09.01	设区的市地方性法规	成都市人大(含常委会)	明确数据来源与收集、加工与存储、删除与转移的合规全程化;不得滥用大数据诱导用户沉迷、过度消费

二、大模型合规监管的现实困境

审视大模型合规监管所面临的现实困境,其主要表现为法律规制体系缺失、监管主体职权不明、应对挑战能力不足、国际规制分歧与合作困境四个方面的问题。

(一)法律规制体系缺失

现有研究表明,大模型在安全治理领域已面临显著挑战。具体而言,大模型应用中易引发决策过程的黑箱化与可解释性缺失、数据隐私泄露及算法歧视等治理难题;在安全维度,更面临恶意利用、模型参数泄露与对抗性攻击等风险。从我国当前监管实践来看,对大模型的规制呈现以网络安全、数据安全及个人信息保护为核心宏观框架,多数规范性文件以面向企业及服务提供平台的倡导性指引为主,整体法律效力层级偏低。在缺乏针对大模型监管的综合性专项法规的背景下,地方层面相关立法亦难以形成有效推进与落地实施的统一路径。综上,我国现行法律规制体系尚未全面回应大模型等新兴技术所衍生的复杂治理需求,亟须构建更具系统性与适应性的监管框架,以保障大模型合规发展。

大模型技术的快速发展已大幅领先于现有法律规范体系的形塑进程,凸显出制度供给与技术演进之间的结构性张力。以欧盟、美国为代表的法域正积极推进大模型监管体系构建,反观我国的推进节奏仍显审慎。2023年7月,我国出台《服务管理暂行办法》,初步对生成式人工智能的训练数据安全与质量等基本问题作出回应。虽然这反映出立法者对人工智能治理法律风险已有初步认知,但尚未针对大模型出台系统性的专门规范。立法者对人工智能大模型治理的审慎立场,彰显了其对公共利益与技术风险的责任担当,但若仅停留于渐进式、被动回应型的制度调适,将难以有效回应该技术在产业落地与社会治理实践中衍生的迫切规制需求。当前,我国大模型法律监管实践中仍面临多重复杂挑战,算法歧视、著作权侵权等核心议题仍未形成系统性解决方案;同时,人工智能技术持续渗透与深度融合正不断重构社会经济系统的运行逻辑,平台化、个性化与去中介化已逐步成为技术应用的主导趋势,对传统产业生态、消费行为模式与劳动力市场结构产生多维度的深刻重构。在此背景下,监管机构亟需构建具备动态适应性的监管框架与规范工具,助力多元市场主体顺畅嵌入持续演化的社会—技术环境,实现技术创新与风险防控的协同推进。

(二)监管主体职权不明

除规范体系不完善外,监管权责界分不清同样制约了合规监管的有效落实。

首先,人工智能大模型监管牵涉网信办、工信部、公安部等多个部门和领域。在现行监管体系下,各部门监管职责划分存在模糊性,易导致监管协调

效率低下、监管职能重叠。其中,网信办的核心职责聚焦于网络信息内容安全与数据安全管理,工信部侧重于网络技术标准制定与产业发展政策规划,公安部则聚焦网络安全保障及网络违法犯罪防治等执法环节。这种职能划分在实践中易引发监管重叠或监管空白的问题。例如,网信办与工信部在互联网数据监管领域的权责界定尚不够清晰,公安部在算法安全监管领域尚未形成体系化的规制策略,且相关部门尚未构建起完善的人工智能风险预警与召回机制。从监管模式看,我国当前网络空间监管呈现多头治理特征,即不同职能部门依据权责划分承担特定领域的监管职责。该模式虽可实现监管领域广泛覆盖,但在跨部门协同治理领域面临显著挑战。以生成式人工智能服务管理为例,网信办在牵头开展监管工作的过程中,仍需进一步强化与其他相关部门的协同联动,以提升监管工作的整体性与协同性。

其次,我国当前针对大模型监管框架采取纵向立法与横向监管措施相结合的双轨推进模式,但现行规范中关键概念界定仍存在模糊之处。例如,《深度合成管理规定》区分了深度合成服务提供者与技术支持者;《服务管理暂行办法》虽明确了生成式人工智能服务提供者的合规义务,却未规定研发者的责任[11];《算法推荐管理规定》以算法推荐服务提供者为规制核心,确立了系统性的合规义务框架。相较于该规定的特定规制范畴,大模型产业的法律关系更复杂——其从技术研发、市场推广到用户应用的全产业链中,法律关系网络覆盖技术研发、技术授权、产品分销等核心环节,错综复杂。进言之,当前对大模型相关概念界定模糊,易引发责任归属认定困境,最终削弱监管问责机制的实践效能。在此背景下,伴随人工智能技术的持续迭代与消费端应用的深度普及,生成式人工智能服务的提供者及技术支持者将面临日益凸显的法律责任压力。因此,当前立法亟须更精确区分与界定上述主体类型,以构建层次分明、覆盖全链条的合规责任体系。

(三)应对挑战能力不足

随着深度学习等人工智能技术快速发展,大模型在架构设计、参数规模及任务处理性能方面持续突破,应用边界也向多领域拓展。但这一进程伴生了一系列新型合规挑战,尤其在数据隐私保护、算法透明性与责任归属认定等核心领域,现行监管体系已暴露出应对大模型合规挑战能力不足的问题。

一方面,模型架构持续演进压缩了监管者对其可解释性和透明度的辨识空间。如 Transformer 架构①的引入,大幅提升了大模型处理复杂序列任务的能力,得益于参数规模的扩大和网络深度的增加,模型能够捕获更复杂的特征模式,进而强化泛化性能[12]。然而,这种性能上的优化以牺牲模型的可解释性和透明度为代价。Transformer 及类似的深度神经网络模型的内部结构和操作极为复杂,宛如一个密不透风的"黑盒"[13],使得外部监管者难以掌握其内部工作原理。

另一方面,监管部门技术手段与法律规范体系的不足,制约大模型监管应对能力的提升。大模型合规性要求其在设计与应用全流程需符合法律法规与伦理准则,而大模型的技术创新常以突破既有规范边界为特征,这一内在属性易引发合规性要求与技术创新之间的张力。具体而言,在用户数据删除请求、知识产权权属争议等典型场景中,企业可能采用未经合规性验证的技术方案予以回应;而监管机构既缺乏标准化的技术验证范式,亦无充足技术能力核验此类方案的实际效能,最终导致合规监管进程滞后于技术实践发展。此外,尽管我国已在《服务管理暂行办法》等立法中初步确立风险分级分类的监管原则,但现行规则多由行业主管部门依行政职权分散制定,尚未在立法层面确立统一、可操作的分级标准与量化区间划分细则。监管体系的技术支撑能力不足,使分级分类监管难以有效落地,无法对大模型在不同风险等级、多元应用场景中引发的差异化挑战形成针对性回应。

(四)国际规制分歧与合作困境

在大模型领域,国际标准统一是实现技术互操作性与推动全球协作的重要基础。以全球开源大模型 DeepSeek 为例,其服务范围覆盖多国用户,但不同法域在数据与隐私保护的立法宗旨、监管框架与执法尺度上存在显著差异,使其在跨境合规实践中面临严峻挑战。一旦未能全面把握特定司法辖区的合规要求,便可能触发法律调查乃至行政处罚。意大利数据保护机构(Garante)因 DeepSeek 未明确告知用户数据收集的具体范围、缺乏合法有效的数据

① Transformer 架构是由 Google 团队于 2017 年在论文 *Attention Is All You Need* 中提出的深度学习模型架构,核心特征是采用"自注意力机制"(Self-Attention Mechanism)。该机制可让模型并行处理序列数据(如文本、语音),无需依赖传统循环神经网络(RNN)的序列化计算,既能高效捕捉序列中不同元素的关联关系,又能支持更大参数规模与更深网络结构的构建,为大模型处理复杂序列任务、捕获细粒度特征模式提供了核心架构支撑。

跨境传输途径,下令将其屏蔽,便为典型案例[14]。这表明,各国对大模型运行机制及隐私风险的认识尚不深入,难以采取精准有效的监管措施,一定程度上阻碍了大模型技术的国际落地与发展。

再者,不同法域针对大模型的监管导向存在明显分歧,进一步加剧了跨国企业合规管理的复杂性与运营成本。欧盟依托《通用数据保护条例》(GDPR)构建起以数据隐私为核心的保护体系,中国则更强调算法透明度与内容合规性要求,这种监管差异可能增加企业跨市场运营的合规成本。虽然国际社会已逐渐意识到协同治理的紧迫性,诸如《全球数据安全倡议》《"中国+中亚五国"数据安全合作倡议》《全球人工智能治理倡议》《首尔宣言》的陆续出台[15],然而受地缘政治、法律传统与产业发展水平等因素影响,各国在大模型及人工智能的监管重点与实现路径上仍存有较大分歧。面对不断加剧的规制风险与制度摩擦,构建适应技术迭代与国际治理需求的大模型法律框架尤为必要。

三、大模型合规监管适应性治理的法治因应

围绕大模型合规监管治理困境的破解,是贯彻落实习近平总书记关于加快人工智能发展和治理的重要论述的应有之义。大模型合规监管适应性治理的法治因应,可从完善法律规范体系、明确主体职权安排、统筹技术能力与合规标准、加强国际交流合作等维度展开。

(一)完善大模型合规监管法律规范体系

鉴于大模型具有技术架构多层次、系统运行复杂及迭代周期快速等核心特征,相关规制体系的设计应立足于大模型技术的开发、部署与应用全生命周期,以保障人工智能技术全流程的有序性与运行效率。从我国人工智能产业发展阶段与市场格局来看,既不适宜直接移植欧盟以严格监管保障用户安全与数据权益,但可能伴随高昂合规成本并抑制产业创新活力的监管路径,也不应盲目效仿美国缺乏统一立法支撑、易引发市场规制碎片化与跨州监管冲突的分散化规制模式,因为这两种模式均可能给我国构建统一有序的市场环境和协同高效的监管体系带来诸多挑战。我国应制定体系化、导向清晰且内容完备的人工智能基本法,通过该法律明确人工智能领域发展的核心目标、基本原则与顶层监管框架,同时为后续下位立法与配套制度的制定提供明确的立法依据与制度衔接基础。

在数据合规层面,首先应遵循透明度原则,对算法开发者、数据控制者及处理者施加明确的责任约束。该约束需涵盖双重核心要求:一是保障训练数据的可追溯性与真实性,建立数据来源核验与全生命周期记录机制;二是对算法关键逻辑进行可解释性说明,以提升用户对算法决策机制的认知与信任。其次,需构建数据分类分级应用与权限管控机制,明确不同敏感等级数据的使用边界,并配套实施周期性安全评估机制,动态监测数据流转中的合规风险。再次,应强化网络平台在内容审核、违规商户处置中的主体责任,通过建立"事前审核—事中监测—事后追溯"的全链条管控体系,构建透明化、可预期、可信赖的数字生态体系,切实保障用户的数据权益与合法利益。最后,可参考欧盟《数字市场法案》的规制思路,通过确立"正面清单+负面清单"的双重规制框架,厘清市场主体的行为边界:借助正面清单界定合法数据流动的场景与路径,通过负面清单禁止数据滥用、不正当竞争等违规行为,推动数据依法有序流动,加强个人数据保护力度,防范市场支配地位滥用风险,维护数字市场的公平竞争秩序[11]。

在算法透明性方面,第一,针对大模型的动态演进特征与技术"黑箱"属性,明确服务提供者的强制性披露义务,要求其完整披露模型的关键信息,包括核心架构、训练数据来源与合规性、决策逻辑的核心机制、任务性能指标及潜在风险类型等方面。第二,需建立模型迭代的全周期记录与说明机制,对迭代过程中涌现的新功能特性与衍生风险进行系统性记录,并由政府监管部门强化模型发布前的前置性技术评估与合规审查,同步推进对第三方评估市场的标准化规范的出台。第三,在算法应用于行政决策、公共服务等涉及公共利益与公众参与的场景中,法律层面除要求算法满足基础透明度标准外,应明确赋予行政相对人及利害关系人算法解释请求权,清晰界定该权利的行使范围、实现路径及救济程序,进而构建算法透明度的实体法与程序法双重保障机制。

在算法反歧视维度,国内大模型虽以中文语料为核心训练数据,但仍存在隐性的价值偏向性问题。为此,需畅通多元主体合作治理路径,推动多元主体协同参与高质量数据集与基准数据库的建设,加快数据标注工具链与标准化体系的研发与落地,以促进人工智能领域数据资源共享体系的规范化建设。围绕算法歧视治理,各类参与主体应在人工智能系统,尤其是在大模型系统的全生命周期内,依法采取有效干预措施来预防、识别并纠正算法歧视,推动算

法决策的公平性实践,保障公众免受非公正算法决策的侵害。

在著作权保护方面,建议通过修订《中华人民共和国著作权法》第 24 条或由最高司法机关出台专门司法解释,明确人工智能模型训练阶段的作品使用行为在合理使用制度中的法律定位,并设置契合比例原则的责任豁免规则,综合考虑使用行为的性质、目的、范围及对著作权人合法权益的影响等因素。与之同时,需进一步明确人工智能生成内容的著作权归属判断标准,以主体对内容表达形式的创造性、贡献程度及身份可确认性为核心,明晰权利归属规则与利益分配路径。

(二)明确大模型合规监管主体职权安排

大模型的合规监管本质上属于系统性治理工程,需依托监管机构、技术开发者、服务提供者、用户及其他利益相关方协同参与联动治理,形成多元主体共建共治的格局。

首先,建议以立法形式明确监管部门的职责边界与协同分工,并确立统一的跨部门协调联动机制。监管机构在政策制定环节需搭建开放式协商平台,主动引入外部智力支持,整合学术界、产业联盟、独立研究机构等多元主体的专业力量,融合跨学科(如计算机科学、法学、伦理学)、跨领域(如技术研发、应用落地、风险防控)的信息资源与多元观点,进而增强监管决策的科学性与合法性基础。以欧盟《人工智能法案》为例,该法案在纵向维度实现了成员国主管机关与欧盟层面监管机构(如欧盟委员会、欧洲数据保护委员会)的协同联动,在横向维度则通过公开征求意见、利益相关方听证会等方式,广泛吸纳行业组织、公众及研究机构参与,最终形成规制知识的多源互补与协同治理格局[16]。此外,可依据人工智能大模型从研发、测试到应用的不同阶段实施分段监管,明确各环节的责任归属主体(如研发阶段的技术提供方、应用阶段的服务运营方)。虽然我国在监管实践方面已经取得了一定进展(如《服务管理暂行办法》的出台),但仍需进一步厘清各部门的职责边界,以破解重复监管与监管空白并存的现实困境。

其次,我国大模型领域专项监管规范体系中,尚未明确大模型研发者的法律概念与权责边界,这一规范空白易引发大模型数据安全风险的责任归属界定模糊问题。从义务主体界定维度来看,欧盟与美国亦未将价值对齐确立为大模型提供者需普遍遵循的核心法律义务,而是通过限定基础模型开发者、政府应用主体等特定范畴,以实现对责任主体范围的精准限定。据此,我国在推进大模型治理体系建设过程中,需先明确区分大模型研发者与大模型服务提供者两类核心主体,并结合二者在大模型全生命周期(研发—应用—运营)不同阶段的角色定位与能力边界,设置差异化的数据安全保障义务体系。具言之,在模型核心训练与调适阶段,研发者需构建覆盖数据采集、预处理至模型训练的技术与管理保障体系,保障模型底层安全,并在模型投入市场前完成系统性安全评估与风险核验;在模型应用服务阶段,服务提供者应通过自行审查或委托第三方专业机构,对数据处理全流程开展合规性实质审查;在模型上线运营阶段,服务提供者还须履行系统实时维护、风险动态预警、突发事件应急响应等持续性运营保障义务,并建立用户数据分级保密机制与全链路安全保护体系。

最后,在大模型多元协同治理体系中,社会公众作为重要参与方,需主动提升人工智能素养,深化对诸如算法偏见、虚假信息生成、数据隐私泄露等大模型技术潜在风险的认识深度,同步强化对技术滥用、误用行为的风险防范意识。与此同时,公众应充分利用反馈机制,向大模型技术开发主体与监管机构输出建设性意见,推动搭建“政府—企业—公众”三方联动的社会协同共治格局,进而为大模型技术的健康、有序发展提供系统性的社会支撑。

(三)统筹大模型的技术能力与合规标准

在人工智能技术范式深度转型的背景下,监管范式亦需同步实现系统性创新,其核心在于构建协同开放的监管权力配置结构、多元融合的监管实施方式、协调一致的监管规制措施三位一体的框架体系。作为适配技术转型的治理方案,包容审慎型治理范式以效率与安全的动态平衡为核心导向:一方面为人工智能新兴业态预留合理发展空间与容错机制,另一方面依据公共风险的动态演化特征实施弹性干预,进而为突破传统规制框架的局限提供制度创新路径。

首先,从技术赋能监管的维度出发,政府需系统提升大模型的技术赋能水平,核心任务在于推进专业化监管平台及配套技术工具的构建与迭代完善。英国在人工智能治理中推行市场化策略,引入独立第三方机构开展人工智能认证服务,服务范围覆盖风险影响评估、合规性审查、算法偏见审计等核心环节,以此保障评估过程的客观性与科学性。美国则聚焦公众参与式治理机制创新,积极探索公众参与式人工智能评估机制的构建路径,依托开放性评估

平台吸纳技术社群与领域专家,对前沿人工智能模型开展治理原则符合性核验与实践应用规范性审查。借鉴国际经验,我国可加快构建社会化人工智能治理服务体系,通过下游专业化标准认证与伦理审查服务,有效承接上游法律法规框架与监管要求,进而显著提升大模型治理的可操作性与实际效能[17]。具体而言,第一,需建设国家级大模型测试验证平台,制订并推广统一的技术标准,明确研发与应用中的安全基准,为行业提供权威技术指引。该平台不仅要提供测试验证服务,还要促进模型供需双方的对接,并具备检测模型可解释性等方面的能力。第二,推动加固工具链的研发与共享,鼓励高质量数据开放,建议由政府主导构建大模型训练与测试数据集,降低数据获取门槛。第三,引入监管沙盒机制,允许企业在限定场景中开展创新试点,依托沙盒的风险隔离特性实现对潜在风险的动态早期识别与精准防控。为规避企业对监管决策潜在干扰,应强化第三方评估机构的独立性与专业性,可参考美国、英国及日本等国设立人工智能安全研究所的经验,推动评估标准与工具的研发与应用[18]。同时,需明确评估机构在专业人员资质、技术工具合规性及资源平台支撑能力等维度的准入标准,并建立动态定期复审机制,确保评估结果的公信力与时效性。

其次,应依据风险等级实施大模型差异化监管。我国当前针对生成式人工智能、深度合成等信息服务虽已提出备案与评估要求,但仍缺乏跨领域风险分级尺度的横向对比机制,且尚未形成统一适用的监管基准。我国人工智能立法需要对分级分类的方式和相关标准予以清晰界定,以实现各领域人工智能主体义务设定的协同适配。可借鉴2024年欧盟《人工智能法案》所采用的"列举+开放"式风险分类框架,构建符合中国国情的风险辨识与评估体系,将分级结果专项纳入人工智能立法。

此外,应积极营造有利于科技创新成果转化的制度环境。可通过监管沙盒机制,在特定区域或场景中为人工智能研发提供风险可控的测试空间,支持模型迭代与优化。同时,设立特定情境下的责任豁免机制,如科学研究、个人学习等非商业性用途,以降低相关主体的合规负担,激发更多主体参与人工智能大模型的探索与创新。在资源支持与促进发展方面,国家应通过财政补贴、税收优惠、政府优先采购等政策工具,加大对人工智能大模型在算力、算法、数据、人才、资本及基础设施方面的投入,有效激发技术创新活力,推动人工智能高质量发展并赋能经济社会转型。

(四)加强大模型生态的国际合作与交流

人工智能大模型的监管治理已成为全球性议题,亟需国际社会构建协同治理机制与制度协作框架。中国在构建契合本土情境的人工智能治理体系时,需深度嵌入全球治理规范网络,与各国在法律框架设计、伦理准则构建、技术标准协同等核心领域开展制度化交流合作。通过批判性借鉴域外前沿治理理念,并将其有机转化为国内监管治理体系的内在构成要素,能够在技术创新与跨国协作间形成正向耦合关系,最终形成双向互促、协同共生的治理格局。

在推进跨国重大人工智能项目落地前,需系统研判主要法域的监管框架与合规要求,将合规管理前置并深度融入本地化运营全流程。首先,通过整体梳理、风险排查与评估业务全链条的合规风险点,动态调整业务策略以最大限度缓释重大项目的合规风险,规避因合规认知缺失或资源投入后遭遇境外监管处罚的风险。在数据治理、算法合规等核心领域,需与本地专业机构或专家充分协作,确保合规判断的本土化适配。其次,应重点剖析不同法域在个人信息跨境流动、科技伦理审查标准及政治文化敏感事项规制等方面的差异化要求,可依托大模型技术设计规制差异知识库,并将其动态反馈至业务决策与执行环节。此外,在与各国监管机构进行沟通互动时,需秉持审慎且积极的沟通策略,充分整合跨境合规律师、技术顾问等外部专业资源,并借助监管机构的答复周期,对监管问询进行细致且周全的回应,塑造专业负责任的企业合规形象,同时可探索监管科技的应用路径,与监管方协同构建技术驱动、常态化运行的合规报送与动态整改机制。

四、结 语

大模型的合规监管是关涉国家发展与安全、个人权益保护的重要议题,备受社会各界关注。本文系统梳理了现行规范性文本在回应大模型发展催生的新型合规挑战时存在的局限性。作为人工智能新质生产力的典型代表,大模型在数据合规、算法透明度、反歧视、著作权保护等多个维度提出了一系列新型合规要求,而现有监管规范与现实发展的脱节进一步加剧了治理张力。究其根本,大模型合规监管仍面临深层次的制度性障碍,亟须以适应性治理进行体系化回应。基于此,宏观上应确立多元协同共治的治理理念,通过构建内在激励机制引导合规行为;微观层面则需聚焦具体问题,提出有针对性的因

应路径,以提升监管的有效性。展望未来,大模型合规监管仍有诸多领域值得深入探讨,如监管科技在合规实践中的具体应用、国际监管规则的协调与互认、伦理原则向法律规则的转化机制等。后续研究可结合实证经验,探索更具前瞻性和系统性的治理路径,推动大模型在合规基础上实现创新与可持续发展。

参考文献:

[1] 网易新闻. 35 家央国企已落地 66 个大模型,国家队引领 AI 变革[EB/OL]. (2024-12-03)[2025-03-07]. https://c.m.163.com/news/a/JIFIHK2P0531WA1P.html.

[2] 苏宇. 大型语言模型的法律风险与治理路径[J]. 法律科学(西北政法大学学报),2024,42(1):76-88.

[3] 张新宝,魏艳伟. 我国人工智能立法基本问题研究[J]. 法制与社会发展,2024,30(6):5-21.

[4] 肖红军,张丽丽. 大模型伦理失范的理论解构与治理创新[J]. 财经问题研究,2024(5):15-32.

[5] 韩旭至. 大模型价值对齐的法治进路[J]. 中国法律评论,2025(1):75-91.

[6] 宋华琳. 人工智能立法中的规制结构设计[J]. 华东政法大学学报,2024,27(5):6-20.

[7] 刘金瑞. 生成式人工智能大模型的新型风险与规制框架[J]. 行政法学研究,2024(2):17-32.

[8] 王正超. 科学数据开放共享中的大模型应用:前景、风险与治理[J]. 现代情报,2025,45(7):167-177.

[9] 应飞虎. 消费者立法中的信息工具[J]. 现代法学,2019,41(2):119-136.

[10] 梁上上. 公司正义:以公司股东的权责配置为视角展开[M]. 北京:法律出版社,2022:17.

[11] 张素华,李凯. 生成式人工智能虚假信息风险与治理研究[J]. 学术探索,2024(7):129-140.

[12] 周辉. 人工智能基础模型安全风险的平台治理[J]. 财经法学,2024(5):3-22.

[13] 郭亚军,李天祥,冯思倩,等. 算法推荐、信息茧房与“附近的消失”[J]. 图书情报知识,2025,42(2):156-166.

[14] 蒋雪颖,刘欣. 开源的代码与隐匿的边疆: DeepSeek 在全球南方舆论场中的数字地缘政治图景[J]. 苏州大学学报(哲学社会科学版),2025,46(4):157-168.

[15] 肖红军,张丽丽. 数字科技伦理国际规则与中国应对[J]. 改革,2024(9):67-83.

[16] 张华平,李林翰,李春锦. ChatGPT 中文性能测评与风险应对[J]. 数据分析与知识发现,2023,7(3):16-25.

[17] 曹建峰. 迈向可信 AI:ChatGPT 类生成式人工智能的治理挑战及应对[J]. 上海政法学院学报(法治论丛),2023,38(4):28-42.

[18] 解志勇. 高风险人工智能的法律界定及规制[J]. 中外法学,2025,37(2):285-305.

(责任编辑:雷彩虹)