人工智能大模型的软法治理及其协调实现

唐士亚
(福州大学 法学院,福建 福州 350108)

摘要:在数字时代,大模型的硬法治理和大模型的涌现式自主生成、企业规模与风险结构差异巨大、训练数据风险隐蔽、技术架构呈分布式等专属特征之间存在系统性冲突,使得硬法治理模式在大模型领域陷入独特的困境。作为一种补充性的治理模式,大模型的软法治理依托自律公约、伦理准则、专业指南、技术标准、公共政策等软法规范,形成引导性、激励性与包容性的治理模式,具有更加尊重大模型市场意志、促进多元主体共治、提供更多治理依据选择和治理针对性强等优势。大模型的软法治理具有复合型的实现路径:一方面,可依靠利益驱动机制、声誉压力机制和技术嵌入机制推动软法规范的功能性实现;另一方面,囿于软法本身的局限性,软法治理应与硬法治理相协调,通过软硬法协同治理以最大限度发挥治理效能。

关键词:人工智能大模型;软法;软法治理;硬法治理;协同治理

中图分类号:D912.1 **文献标志码**:A **文章编号**:2096-028X(2026)02-0129-08

以 OpenAI GPT-4、Google Gemini、DeepSeek 等为代表的人工智能大模型(简称大模型)给人类社会带来了深刻影响:一方面极大提升了生产力与生产关系的智能化水平,另一方面又引发了人工智能技术应用的风险隐忧与治理博弈。目前,中国总体上以国家互联网信息办公室(简称网信办)为牵头部门,联合工业与信息化部、国家发展与改革委员会、科技部、公安部等开展对大模型行业的协同治理。其中,网信办承担主要的牵头监管职能,省级网信部门负责属地内生成式人工智能应用的登记工作。在治理依据上,2023 年 8 月 15 日正式施行的《生成式人工智能服务管理暂行办法》是针对大模型的主要监管规范,对生成式人工智能的服务规范、数据与算法治理、用户权益保护及分类分级监管等作出了详细规定。在配套法律法规上,《中华人民共和国网络安全法》(简称《网络安全法》)、《中华人民共和国数据安全法》(简称《数据安全法》)、《中华人民共和国个人信息保护法》(简称《个人信息保护法》)、《互联网信息服务深度合成管理规定》、《互联网信息服务算法推荐管理规定》等法律法规也是大模型监管的重要依据。上述法律法规依托国家强制力保障实施,具有刚性约束力与明确制裁机制,属于硬法的范畴。① 这种依托行政监管部门落实,以具有国家强制力的法律法规、部门规章等硬法规范为依据的治理方式,本文称之为大模型的硬法治理模式。

与硬法和硬法治理模式相比,软法作为不依赖国家强制力但有事实上效力的行为规范,是对硬法规范代表的公力执行机制的替代与补充,②以私人执行的方式为公共治理提供一种更低成本、更具灵活性的行为选择。因此,大模型的软法治理模式是指依据不具有国家强制力但有实际引导与规范作用的公共政策、伦理准则、技术标准和自律公约等软法规范构建的针对大模型的开发、部署与应用的一种柔性治理模式。

本文针对当前大模型硬法治理的困境,对事实上已经存在但缺乏系统性提炼的大模型软法治理模式进行梳理与阐释,重点论证大模型软法治理的原理与作用机制。需要指出的是,大模型的软法治理和硬法治理不是冲突对立的关系,而是共同构成“硬法为主,软法补充”的复合结构。在动态演进的技术环境中,软硬法的协同治理模式将会是大模型治理的长期趋势。

基金项目:2023 年度教育部人文社会科学研究规划基金项目“金融科技的软法治理及其协同问题研究”(23YJA820026)
作者简介:唐士亚,男,福州大学法学院副教授、硕士生导师。

① 参见沈岿:《自治、国家强制与软法——软法的形式和边界再探》,载《法学家》2023 年第 4 期,第 40 页。
② 参见罗豪才、周强:《软法研究的多维思考》,载《中国法学》2013 年第 5 期,第 104 页。

一、人工智能大模型硬法治理的困境

大模型具有参数规模庞大、海量训练数据、存在涌现能力及具备多任务学习能力等特点，能够在自然语言处理、图像识别及语音识别等众多领域得到广泛应用。大模型的硬法治理除了面临监管法律滞后、监管一刀切、社会共治缺位等数字市场治理的共性问题之外，还因为硬法治理的刚性、职权性、分立性等规范特点，和大模型的涌现式自主生成、企业规模与风险结构差异性大、训练数据风险隐蔽及技术架构呈分布式等专属特征之间存在系统性冲突，使得硬法治理模式在大模型领域呈现出独特的困境。

（一）大模型硬法治理的责任框架与大模型涌现式自主生成相割裂

硬法治理的责任框架以可预见的法律行为、可预测的行为后果、主观过错为核心要件，责任的归责逻辑高度依赖“行为—结果”的因果关系。但大模型具有独特的涌现式自主生成特性，与硬法治理的责任原理具有明显区别，构成了大模型治理的特殊困境。具体而言，大模型依托深度神经网络对数百万乃至数十亿的数据进行训练，其推理与运算过程非常复杂，能够在脱离开发者预设程序与人工干预的情况下，自主生成全新、非预设、随机性的内容与决策，具有“技术黑箱”特征且缺乏可供人类理解的直观解释。① 硬法治理以故意、过失为核心的主观归责要件，无法适配大模型无独立主观意志却可自主衍生风险的特殊形态，既不能将模型生成的有害结果直接归责于大模型的主观过错，也无适配的法定责任兜底规则。

硬法的责任框架追求责任结果的确定性、唯一性及标准性，而大模型基于概率算法运行，输出结果具有不确定性、随机性和语境依赖性，同一指令会产生差异化输出结果。此外，大模型具有的“幻觉”风险、偏见风险及虚假信息风险等，其触发亦无固定规律可循。② 这使得刚性的硬法归责规则难以与概率化的大模型运行模式相匹配，二者的治理结构形成天然错位。

（二）大模型硬法治理的标准与企业规模结构及风险结构不匹配

大模型市场中存在着规模不等的研发企业，各类企业在研发水平、资金规模等方面存在巨大差异。大模型分为基础大模型、行业大模型、小型专用模型等类型，不同类型的大模型在风险特征方面呈现质的差异。当前大模型硬法治理标准采取“一刀切”的做法，既未充分考量大模型的企业规模结构差异，也与大模型的风险结构不匹配。

首先，不同规模的企业在大模型的研发逻辑方面存在巨大差异。大型科技企业拥有雄厚的资金实力、丰富的技术人才储备和强大的研发能力，能够投入大量资源进行大模型的深度研发和复杂应用场景的探索。而中小型企业则在资源和能力上相对薄弱，更多地专注于一些特定领域的小型模型研发或大模型的应用拓展。然而，现有的硬法治理往往简单地采用统一的标准进行监管。例如，在大模型的算法透明度标准方面，无论企业规模大小，都被要求以同样详尽的程度公开算法逻辑与决策过程。这对于中小型研发企业而言，无疑是巨大的研发成本与合规成本负担。因为要达到与大型企业同样的算法透明度标准，中小型研发企业的成本负担将会极大增加，由此有可能直接导致中小型企业大模型业务的萎缩乃至退出市场。③

其次，基础大模型、行业大模型和小型专用模型在风险结构等方面也存在显著差异。三者的风险等级、影响边界与外溢效应完全不同，导致统一监管标准天然存在适配性缺陷。基础大模型通用性强、参数规模庞大，风险具备全局性、基础性与强外溢性，风险事件易向全行业传导扩散，理应进行更严格的准入、评估与合规约束。行业大模型聚焦垂直领域，风险集中在特定行业场景，仅产生领域内业务关联风险，无需套用基础大模型的严苛监管规则。④ 小型专用模型参数规模小、场景单一，以局部碎片化风险为主，影响范围有限、外溢性极弱。若采取“一刀切”的统一监管标准，一方面无视三类模型风险特质的分层差异，违背风险与监管强度相匹配的治理逻辑；另一方面会造成监管资源错配，要么对低风险的小型模型施加过度监管、增加合规成本，要么无法对高风险的基础大模型形成有效的从严约束。

---

① 参见姚佳：《人工智能的训练数据制度——以“智能涌现”为观察视角》，载《贵州社会科学》2024年第2期，第52页。

② 参见何静、沈阳、谢润锋：《大语言模型幻觉现象的分类识别与优化研究》，载《计算机科学与探索》2025年第5期，第1297页。

③ 参见赵精武、文彬：《论人工智能技术治理体系的中小企业特别治理机制》，载《湖北社会科学》2024年第10期，第126页。

④ 参见肖静华、盛君叶、罗婷予：《行业大模型的构建：从企业知识创造到行业知识积累》，载《经济纵横》2026年第1期，第72页。

(三)大模型硬法治理的内容审查模式滞后于训练数据的隐性风险

当前中国针对大模型生成内容的硬法治理体系,在治理逻辑上沿袭传统互联网内容的监管模式,治理重心集中在大模型对外输出的文本、图像等内容,重点排查违法违规信息及不良有害信息等可直接识别的显性问题。① 而大模型的内容输出依托海量训练数据,其中所蕴含的价值偏差、认知偏见和虚假内容等问题,并非是在内容生成环节偶然产生的,而是源于训练数据中隐藏的价值歧视、碎片化虚假信息及不良导向等。这些训练数据的隐性风险分散于海量的原始数据之中,经过大模型算法的深度学习和逻辑拟合后,被内化为大模型的底层能力和知识体系,无法通过事后内容审查进行识别、溯源和整改。② 大模型硬法治理尚未有训练数据合规筛查的法定标准,也缺乏稳定的针对数据偏见、数据虚假及数据污染等问题的治理机制。实践中主要依靠事后追责、内容下架等末端治理手段,只能解决表层的内容生成风险,而无法从源头上解决大模型和训练数据的内部问题,导致大模型硬法治理陷入“治标不治本”的困境。

(四)大模型硬法治理的地域管辖边界与大模型技术架构存在冲突

大模型硬法治理以物理空间为基础划定地域管辖边界,依托属地和属人管辖原则确立较为清晰的监管范围与职责。大模型天然具有分布式部署、算力跨境调度和训练数据跨境流动等技术架构特征,这与大模型硬法治理的管辖方式构成了明显冲突。具言之,大模型的训练数据来源于互联网上的诸多国家与地区,模型参数分布于不同法域的算力集群之中,内容输出服务通过互联网向全球用户提供,突破了传统法律以行为地、结果地划定管辖连接点的适用基础。基于数据主权原则的大模型硬法治理规范多设定了属地管辖、数据本地化存储、数据出境审查等强制性规则,但大模型具有的训练数据跨域采集、内容输出全球化等技术特性,使单一法域管辖难以覆盖全大模型的全流程运营,极易出现管辖重叠、监管真空与规则竞合问题。③

蝴蝶效应旗下“Manus AI 跨境并购案”的管辖权冲突即为典型例证。Manus 由中国团队主导研发,训练数据整合全球网络资源,核心算力集群部署于中国境内,模型参数则分布式存储于新加坡、美国等不同法域的算力设施,服务端口面向全球开放。为规避监管,Manus 于 2025 年 6 月将总部迁至新加坡,大幅裁撤国内团队、停止境内服务,试图通过变更注册地实现“去中国化”,并于同年 12 月以约 20 亿美元对价被美国 Meta 公司收购。2026 年 4 月 27 日,国家发改委外商投资安全审查工作机制办公室依法作出禁止投资决定,要求撤销整笔交易,这是《外商投资安全审查办法》实施以来,中国首例公开叫停的 AI 领域外资收购案。④ 在该案中,形成了中国基于技术本源与数据安全的属地监管、美国基于外资审查的域外管辖、新加坡基于注册地的形式管辖之间的叠加冲突,印证了大模型技术架构与硬法治理的地域管辖逻辑之间的现实矛盾。

二、一种补充性的治理进路:人工智能大模型软法治理及其原理阐释

(一)人工智能大模型软法治理的相对优势

第一,软法治理更加尊重大模型市场意志,创设包容性治理空间。大模型具有技术迭代快、应用场景丰富、创新性和不确定性强等特点,单纯依赖硬法治理容易出现治理滞后或过度治理的问题,抑制市场活力。以大模型技术创新场景为例,行业协会等软法制定主体通常采用“引导+建议”的柔性治理模式,依托行业公约、自律规范、技术指南等软法工具,为市场主体提供行为指引而非刚性约束。⑤ 这种模式充分尊重大模型企业在算法研发、数据训练和商业化探索中的自主决策空间,通过行业共识、声誉约束与自我合规承诺实现大模型的风险管控。

第二,软法治理有助于激励大模型市场中公私多元主体参与合作治理。这种合作治理的形成原理在于,软法所具有的引导性与共赢性特质,有助于增强公私主体间的互信基础,进而降低合作协商成本,最终推动

---

① 参见许立勇、高宏存:《“包容性”新治理:互联网文化内容管理及规制》,载《深圳大学学报(人文社会科学版)》2019 年第 2 期,第 53 页。

② 参见黄锫:《人工智能大模型训练数据的风险类型与法律规制》,载《政法论丛》2025 年第 1 期,第 26 页。

③ 参见高奇琦:《主体间性视域下的大模型共识性治理》,载《山西大学学报(哲学社会科学版)》2025 年第 1 期,第 53 页。

④ 参见佘惠敏:《AI 出海不可触碰监管红线》,载《经济日报》2026 年 5 月 4 日,第 3 版。

⑤ 例如,中国互联网协会发布的《人工智能通用大模型合规管理体系指南》(T/ISC 0080—2025)作为团体标准,仅提供合规管理的建议性框架与最佳实践。中国证券投资基金业协会发布的《基金经营机构大模型技术应用规范》(T/AMAC 0004—2026)同样以技术与管理建议为主,附录提供风险应对建议与应用案例,体现自律导向。

公私主体从“对抗监管”转向“协同共治”。[①] 相较于硬法的刚性约束,软法更强调沟通、共识与激励相容,能够有效吸纳大模型研发企业、行业组织、科研机构等市场主体的专业知识与实践经验,减少监管部门与监管对象之间的信息不对称。例如,监管部门可通过颁布促进型公共政策,明确大模型行业的发展方向及重点领域,引导大模型企业凝聚行业共识并形成稳定的未来预期。以地方政府的相关公共政策为例,《深圳市打造人工智能先锋城市的若干措施》提出在政策有效期内,每年发放最高5亿元“训力券”,降低人工智能模型研发和训练成本;每年发放最高1亿元“模型券”,降低人工智能模型应用成本,[②]直接形成对当地大模型企业的有效激励。

第三,大模型软法治理具有针对性强、响应及时和预防性治理的相对优势。软法治理无须遵循严格、繁琐的立法程序,制定主体灵活多元,能够紧密跟踪大模型技术演进与市场供需变化,针对内容安全、数据合规和算法偏见等具体问题快速形成规则供给。无论是自律公约、技术标准、伦理准则还是企业合规指引,都有助于在大模型各类风险尚处于萌芽滋生阶段便完成前置性引导与预防性治理,实现治理关口主动前移,达成大模型治理领域“治未病”的成效,有效克服硬法治理存在的滞后性与粗放性。

(二)人工智能大模型软法治理中的软法规范

在大模型治理领域,涉及的典型软法规范主要有自律公约、伦理准则、专业指南、技术标准和公共政策等。[③] 大模型的不同类型软法规范在适用中具有事实上的位阶差异,且不可避免地存在着规范之间的冲突问题。下文将从类型梳理、位阶原理、冲突整合这三个方面,对大模型软法规范进行一个统合性描述。

1.大模型软法规范的类型

自律公约、伦理准则、专业指南属于典型的软法规范。[④] 中国已出台具有代表性的大模型行业自律公约,如2024年8月发布的《生成式人工智能行业自律倡议》。[⑤] 该自律倡议围绕数据和算法模型安全合规、促进内容生态建设、遵循价值观与伦理道德标准等提出了自律准则,但也存在原则性强、细则不足、可操作性弱的不足。《政务领域人工智能大模型部署应用指引》于2025年10月发布,是典型的大模型专业指南,为各级政务部门提供人工智能大模型部署应用的工作导向和基本参照。[⑥] 该指引是中国首个国家层面针对政务大模型的系统性指引文件,但也存在原则性表述多、量化细则与标准缺失的不足,容易导致基层执行过程中的标准不一、自由裁量权过大等问题。

在大模型的技术标准中,需要对强制性国家标准和推荐性国家标准进行区分。大模型企业必须遵守强制性国家标准,一旦违反则需要承担相应的法律责任,故强制性国家标准具有硬法属性。推荐性国家标准的效力是鼓励采用、自愿遵守的,因此推荐性国家标准才属于实质意义上的软法规范。中国大模型行业典型的推荐性国家标准有《人工智能 面向机器学习的数据标注规程》(GB/T 42755—2023)、《人工智能 服务器系统性能测试方法》(GB/T 45087—2024)、《人工智能 机器学习系统技术要求》(GB/T 43782—2024)等。

大模型行业的公共政策是政府部门为引导大模型行业健康发展、规范行业秩序、推动技术落地应用而制定并发布的一系列具有行政指导力的意见、办法与行动计划。中国大模型行业具有代表性的公共政策有2025年8月印发的《国务院关于深入实施“人工智能+”行动的意见》等。需要指出的是,公共政策之所以多属于软法,是因为大多数公共政策以规划、纲要、指导意见等非规范性文件形式存在,缺乏严格的法律约束力与明确的法律责任,在实施机制上更多依赖引导、协商和激励等柔性手段。但并非所有的公共政策都属于软法,因为部分公共政策会通过法定程序转化为法律法规,具备刚性拘束力与国家强制执行力,落入硬法规范范畴;还有部分仅属于临时行政安排、事务性工作部署,缺乏普遍约束力与反复适用效力,不满足软法的规范

① 参见柳亦博:《论合作治理的路径建构》,载《行政论坛》2016年第1期,第13页。

② 参见《深圳市打造人工智能先锋城市的若干措施》,载深圳政府在线2025年1月2日,http://www.sz.gov.cn/cn/zjsz/fwts_1_3/tzfw/yhzc_1/content/post_11934722.html。

③ See Gary E. Marchant & Carlos Ignacio Gutierrez, *Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence*, Minnesota Journal of Law, Science and Technology, Vol.24:375, p.384-385(2023).

④ 参见梁剑兵、张新华:《软法的一般原理》,法律出版社2012年版,第63页。

⑤ 参见《2024年中国网络文明大会人工智能论坛在成都举行》,载中央网络安全和信息化委员会办公室、中华人民共和国国家互联网信息办公室网站2024年8月29日,https://www.cac.gov.cn/2024-08/29/c_1726614903460429.htm。

⑥ 参见《政务领域人工智能大模型部署应用指引》,载中央网络安全和信息化委员会办公室、中华人民共和国国家互联网信息办公室网站2025年10月10日,https://www.cac.gov.cn/2025-10/10/c_1761819469929310.htm。

构成要件。

除此之外,在大模型治理领域,还存在着合规指引、行业倡议、示范合同、技术白皮书和管理体系规范等软法规范。这些软法规范与前文详细说明的软法规范在适用原理上具有相似性,本文不再详细展开说明。

2.大模型软法规范的位阶

大模型软法规范的位阶,是指在大模型软法体系内部,不同软法规范之间依据制定主体、适用范围、效力强度等形成的相对效力等级与优先顺序。① 不同于硬法等级鲜明的层级体系,大模型软法规范的位阶具有相对化、自治化和事实性的特点,且整体位阶低于硬法,不得与硬法相抵触。在此基础上,大模型软法规范的位阶顺序,可依据"制定主体层级越高、适用范围越广、效力越接近硬法的软法具有更高位阶"的原则来进行相对效力等级与优先顺序的排列。

具体而言,大模型推荐性国家标准、国务院及有关部委制定的大模型行业政策等依托国家权威制定,适用范围覆盖全国,对行政监管、大模型市场具有较强的实际约束力和指导效力,效力位阶最高。其次为各省市等地方层面出台的大模型推荐性地方标准、区域治理政策与地方政府指引性文件等,仅适用于特定行政区域,不得违背硬法与高位阶软法规范。效力位阶更低的是社会自治类软法,包括大模型行业协会制定的自律准则、团体标准等,依托市场与行业自律实施,无国家强制力约束,仅对行业成员与市场主体产生约束效力。

3.大模型软法规范的整合

大模型治理中的软法规范具有软约束性,可以理解为各方基于"最低限度的目的"所达成的共识性妥协安排。② 这也造成了软法规范在内容和效力上的不确定性,即一些软法条文不够精确,对一个条款可能作出不同解释而导致结果或效力的不确定性。同时,在大模型产业发展进程中,伦理准则、自律公约和公共政策等软法规范在价值导向上存在立场差异,导致软法规范内部的现实冲突。例如,大模型的伦理准则多由高校、研发企业制定,侧重于价值引领与道德约束;行业自律公约以产业发展和商业实践为核心,注重实操性和效益性;各地政府出台的大模型产业政策立足于地方经济发展和属地管理需要,强调风险管控前提下的大模型产业发展。因此,需要适当引入体系性方法对大模型治理中的软法规范进行体系性整合。

具体而言,一是凝聚各方对大模型软法治理的共识。可建立常态化多方协商平台,汇聚企业、用户、研究机构及公众代表,共同界定核心治理原则,通过发布联合倡议、制定行业公约及最佳实践指南等软法规范,将抽象目标转化为共识性软法文件。二是完善软法实施的监督与反馈闭环。搭建公开透明的软法落实情况监督渠道,由企业、行业协会、社交媒体与用户共同监督软法规范执行效果,构建畅通的投诉、反馈、评议与整改回访流程,以公共监督强化软法约束力,避免软法流于形式。三是推动跨领域软法规范的衔接协同,比如技术标准与伦理准则的配套、行业公约与企业内部规范的呼应,明确同一行为在不同软法规范下的一致要求,从而提升软法规范的系统性与可操作性。

(三)人工智能大模型软法治理中的自治与共治

大模型的软法自治主要由与大模型研发与应用相关的企业、行业组织等主体主动推进,依据自身技术特性、业务场景及发展需求进行自我规制,规制对象包括"平台行为"和"用户行为"两个方面。③ 在实践中,不少大模型企业会专门设立内部伦理委员会,同时在具体业务部门配置伦理专员或专项工作小组,确保人工智能伦理准则能贯穿于大模型技术迭代、产品落地的每一个环节。例如,商汤科技于 2020 年 1 月成立人工智能伦理与治理委员会,由内外部委员共同组成,下设有秘书处、执行工作组和顾问委员会,从算法、数据及社会影响三个方面对商汤所有产品线进行审核。④

大模型的软法共治是指通过多元主体合作共治,使企业、消费者和行业协会等都能够作为主体参与大模型治理,而非作为被动的被监管者。在共治模式中,不同主体承担着差异化且互补的角色。监管部门无需过度介入技术细节,而是通过出台公共政策、发布治理原则与指导指南,为大模型的发展划定合法合规的框架。

① 参见王锴:《法律位阶判断标准的反思与运用》,载《中国法学》2022 年第 2 期,第 7 页。

② 参见张耀元:《国际经贸条约治理新范式:软约束规则的兴起与实施》,载《环球法律评论》2025 年第 6 期,第 210 页。

③ 参见邓栩健:《生成式人工智能平台自治规则的体系化构建》,载《东方法学》2025 年第 3 期,第 81 页。

④ 参见《商汤科技人工智能伦理与治理委员会简介》,载商汤科技网站 2024 年 10 月 23 日,https://www.sensetime.com/cn/ethics-detail/56847?categoryId=32429。

行业协会凭借其第三方中立属性,可针对企业的自治情况、政府的政策落实效果开展独立监督与评估,比如核查大模型企业伦理制度的实际执行情况、评估大模型应用对社会公共利益的影响。公众作为大模型技术的直接使用者与潜在影响对象,能够通过产品反馈、意见征集和社会监督等渠道,将自身的利益诉求、使用体验传递给治理主体,推动治理决策更贴合民生需求。

大模型软法治理中的自治与共治并非相互独立,而是呈现出深度融合、相辅相成的紧密关系。① 一方面,自治是共治有效开展的基础前提:若大模型企业缺乏基本的自治意识,未建立完善的内部伦理治理机制,行业协会也未形成统一的自律规范,那么政府制定的公共政策将难以落地,公众的诉求也无法通过合理渠道传递,整个共治体系便会陷入"无的放矢"的困境。另一方面,共治是自治持续深化的重要保障:政府通过政策引导与资源支持,鼓励大模型企业加大对自治体系的投入,比如对建立完善伦理治理机制的企业给予政策优惠或项目支持;社会组织的监督评估可倒逼企业不断优化自治措施,避免企业因追求短期利益而忽视技术风险。②

三、人工智能大模型软法治理的实现路径

(一)人工智能大模型软法治理的内部路径

一是利益驱动机制。利益驱动机制可通过权利义务驱动、成本收益驱动和资格荣誉驱动三个方面加以实现。首先,在大模型治理中,通过明确权利、义务与责任的匹配关系以激发大模型企业的合规动力。对于主动落实行业自律公约、伦理准则的大模型企业,监管部门可建立合规激励清单:允许其优先接入国家级公共算力平台与训练资源池,享受算力配额倾斜与成本优惠;在伦理审查环节实行简化流程、缩短时限和差异化备案等便利措施,适度减轻合规成本。其次,政府可对合规记录良好的大模型企业给予研发补贴、优先采购产品等经济激励,对大模型企业的数据滥用等违规行为处以罚款,使违规收益远低于合规收益,形成成本、收益的正向导向。最后,政府与行业协会可开展行业荣誉评选,如开展"人工智能伦理应用示范企业"等评选活动,通过官方媒体、行业协会宣传表彰,提升合规大模型企业的品牌价值与社会认可度。

二是声誉压力机制。企业声誉作为一种信号,体现了消费者和社会公众对企业的认知程度和信任程度。因此,声誉的建立和维护就成为一种不需要外界强制而能够自我实施的过程,这就是声誉机制的基本原理。③ 大模型治理中的声誉压力机制,其实质在于将伦理准则、自律公约等软法规范通过社会舆论、行业评价及消费者认知转化为可感知的声誉收益与成本,引导大模型企业主动合规经营。换言之,声誉压力机制也是软法规范以企业声誉为中介,实现软法的间接性约束效力的过程。例如,当大模型存在数据偏见、虚假信息传播等违法违规问题时,公众质疑与媒体曝光会直接损害其品牌声誉与市场信任度,进而影响用户选择、投资意愿及政策支持。反之,若大模型在公平性、透明性、安全性等方面表现优异而具有良好声誉,开发者将获得社会认可、商业合作机会与政策激励。这种以声誉奖惩为核心的非强制性约束机制,形成"合规增信、违规失信"的动态约束,推动大模型企业主动践行"技术向善"。

三是技术嵌入机制。技术嵌入机制是指将伦理准则、技术标准和自律公约等软法规范通过技术手段嵌入大模型研发、训练、部署和迭代的全生命周期,从而通过算法代码实现对于大模型的软法约束,确保大模型的价值对齐。具体来说,在模型训练阶段,嵌入偏差惩罚项以抑制数据偏见,从而落实大模型伦理准则中的公平性原则;在模型部署与应用阶段,嵌入隐私计算、算法影响评估和内容过滤等技术标准,保护用户隐私和个人信息安全,契合大模型的安全性、透明性准则。由此可见,技术嵌入机制本质上是通过技术手段将软法规范转化为大模型的默认行为,既降低了人为干预的不确定性,也推动软法落地。

(二)人工智能大模型的软硬法协同治理路径

1.大模型治理中的软硬法规范交互协同

硬法与软法作为各自独立的法范畴,不存在从属依附关系,而是并行不悖地共存于法律规范体系之中,

① 参见高奇琦:《共有、共治与共享:大模型技术的发展之路》,载《华中科技大学学报(社会科学版)》2024年第2期,第61页。
② 参见唐士亚、张巍瑜:《金融科技伦理规制的基本法理与制度化构造》,载《北京邮电大学学报(社会科学版)》2022年第6期,第25页。
③ 参见雷宇:《声誉机制的信任基础:危机与重建》,载《管理评论》2016年第8期,第226页。

并在不同场域中发挥约束力,或在同一领域中运用不同方式或侧重不同方面来规范社会关系。① 为实现大模型软硬法的协同治理,需要将二者统合于同一治理框架之中,厘清二者之间的交互逻辑。

一是软硬法的内容互补。硬法因立法程序严谨、修订周期长,往往聚焦于原则性框架与底线规则,例如通过《网络安全法》《数据安全法》《个人信息保护法》确立网络主权和算法公平等基础原则,为模型开发划定不可逾越的红线。而软法则以技术标准、行业指南及伦理公约等形式,对硬法难以触及的细节领域进行填补。例如,针对生成式 AI 的“幻觉”问题,软法可制定大模型输出内容的可验证性技术标准,要求大模型标注信息来源、声明置信度;面对算法歧视风险,行业协会可发布算法公平性评估指南。这种“硬法定原则、软法定细则”的互补结构,既避免硬法因过度细化而束缚创新,又防止软法因缺乏强制力支撑而流于形式。

二是软硬法的体系转化。在大模型治理中,硬法可对实施效果良好且运行成熟的软法内容加以吸纳,将其转化为法律规范的内容。例如,开源社区形成的大模型安全审计标准,若经实践验证可有效降低技术滥用风险,后续可被转化为人工智能专项立法的强制性认证条款;企业制定的数据隐私保护自律承诺,若成为行业通行惯例,亦可被立法机关吸纳,进而上升为法定义务。反之,硬法条款也可以被转化为具有实操性的软法规范。由于硬法不可能进行全面规范,因此需要较为灵活的软法明确相关事项,从而形成治理的全貌。② 例如,《数据安全法》《个人信息保护法》中关于数据安全和个人信息保护的条款难以直接适用于大模型训练数据核验和内容输出审核等具体场景。对此,《网络安全技术 生成式人工智能服务安全基本要求》《网络安全技术 人工智能生成合成内容标识方法》《翻译行业生成式人工智能应用指南(2025)》等具有软法性质的国家标准和指南性文件,将硬法条款转化为训练数据来源与标注、模型生成内容技术要求等软法规则。

三是软硬法的适用竞争。软硬法的适用竞争主要取决于经济指标上的交易成本因素及道德指标上的民主性或正当性因素。③ 在多数情况下,硬法在二者的适用竞争中处于主导性地位,而软法通过非对抗性的柔性治理方式,能在很大程度上协调各方利益、降低治理成本,从而在某些特定场景中成为更优选择。因此,软硬法的适用竞争会依据场景特性形成主次分明的治理格局。在国家安全、公共利益等核心领域,硬法凭借国家强制力占据主导地位,例如对深度伪造技术的监管、关键信息基础设施的算法备案等,均需通过法律明确责任主体与处罚措施。而在算法优化、内容生态治理等快速迭代领域,软法则通过市场机制发挥灵活性优势,如平台通过用户协议约束大模型滥用行为、行业协会制定内容审核标准,既能快速响应技术变化,又可通过用户选择形成“用脚投票”的约束效应。

2.大模型治理中的公私组织协同共治

一是私主体主导型共治。私主体主导型共治是指在大模型治理中,以企业、科研机构等私主体为核心展开的协同共治模式。私主体凭借自身在技术研发、数据处理和市场运营等方面的优势,在大模型的开发、部署与应用环节发挥关键作用。他们能够快速响应市场需求和技术变革,灵活调整治理策略。同时,私主体之间通过共同制定代码审计标准、安全测试标准与伦理准则,形成自律机制。在这种模式下,监管部门则主要负责设定监管底线与法律红线,比如明确训练数据的使用边界和侵权处罚措施,以此引导大模型企业在合规基础上进行创新,平衡技术发展、商业效益与公共利益之间的关系,推动产业有序发展。

二是公主体主导型共治。公主体主导型共治以监管部门为统筹核心,运用标准、指引等软法工具搭建治理框架,从而引导企业、研发机构等私主体规范自身行为,系统性管控大模型的潜在风险。其治理逻辑体现为“后设规制”理念,即政府减少具体监管,转而通过政策引导、标准制定与激励约束塑造治理环境。④ 具言之,政府无须直接制定刚性细则干预大模型市场,而应转向框架性引导与间接治理。政府先确立统一的治理底线、价值原则与合规评价标准,将具体行为规范、伦理约束与自律细则交由行业协会、大模型企业等市场主体。政府则以监督者身份对市场自治规则进行审查、评估与纠偏,核查大模型市场的自治体系是否达标,督

① 参见王兰:《民间金融软法治理及协同问题研究》,厦门大学出版社 2024 年版,第 139 页。

② 参见邱静:《人工智能领域中软法治理的国际实践与中国思索》,载《西华大学学报(哲学社会科学版)》2026 年第 1 期,第 24 页。

③ 参见罗豪才、宋功德:《软法亦法——公共治理呼唤软法之治》,法律出版社 2009 年版,第 399 页。

④ 后设规制也常译为“元规制”,其本质是“规制的规制”,强调监管者与被规制者之间的互动性和灵活性,适用于技术迭代快、行业复杂度高、传统“命令—控制型”规制难以全面覆盖的领域(如金融科技、人工智能等前沿领域)。

促大模型软法规范贴合公共利益与社会伦理。①

在实践中,基于“后设规制”的大模型公私组织协同共治可以从训练数据、算法机制和内容审查等方面展开。在训练数据中,监管部门确立数据合规、隐私保护和内容安全的底线框架,由行业协会出台数据筛选、语料清洗和版权核验的技术指引,督促企业自主规范数据采集、训练全流程,防范数据侵权与偏见风险。在算法机制中,监管部门划定算法公平、透明、可控的核心原则,依托行业自律规范引导企业建立算法溯源、风险检测和偏差修正机制,自主优化模型迭代与运行逻辑,解决算法歧视、过度拟合等问题。在内容审查中,监管部门明确生成内容的合规标准与伦理标准,授权行业协会搭建统一的内容治理参考标准,由企业落实事前筛查、事中监测、事后溯源的自律管理。

四、结语

在数字时代,大模型的广泛部署在推动新质生产力发展的同时,也带来了数据安全保护与产业治理方面的问题,即需要在“发展与治理”“安全与效益”等方面取得更加合理的平衡。强调回应性、协商性、合作性的软法治理为大模型提供了一种更加包容开放的治理方案,也打破了政府、监管部门是治理资源唯一合法来源的僵化思维。软法治理通过利益驱动机制、声誉压力机制和技术嵌入机制三条路径,有效促成了大模型治理不同主体间的利益协调与资源共享,丰富了大模型治理的工具箱。当然,单一的软法治理无法全面防范大模型固有风险,需要通过软硬法规范交互协同和公私组织协同共治的方式,才能真正实现对大模型治理的修正与补全。展望未来,软法治理将会在人工智能时代占据愈发重要的位置,软法在人工智能治理中的实施主体、正当程序、规范运行和效力边界等问题,亟须进一步厘清。

Soft Law Governance of Large-Scale AI Models and Its Collaborative Realization

TANG Shiya

(Law School, Fuzhou University, Fuzhou 350108, China)

Abstract: In the digital age, there are systematic conflicts between the hard law governance of large-scale AI models and their emergent autonomous generation, significant differences in enterprise scale and risk structure, the concealment of training data risks, and the distributed technical architecture, etc. This has led to a unique predicament for the hard law governance model in the field of large-scale AI models. As a supplementary governance model, the soft law governance of large-scale AI models relies on soft law norms such as self-discipline conventions, ethical guidelines, professional guidelines, technical standards, and public policies to form a guiding, incentive, and inclusive governance model. It has the advantages of respecting the will of the large-scale AI model market, promoting multi-party co-governance, providing more governance basis choices, and having strong targeted governance. The soft law governance of large-scale AI models has a composite implementation path: on the one hand, it can rely on interest driven mechanisms, reputation pressure mechanisms, and technical embedding mechanisms to promote the functional realization of soft law norms; on the other hand, due to the limitations of soft law itself, soft law governance should be coordinated with hard law governance, so as to maximize the governance efficiency through the collaborative governance of soft and hard laws.

Key words: large-scale AI models; soft law; soft law governance; hard law governance; collaborative governance

(责任编辑:盛 怡)

① 参见王秋艳、万国华:《生成式人工智能的双重法律规制理路》,载《中国特色社会主义研究》2023年第4期,第77页。