

数字政府建设中多轮蒸馏大模型应用的行政法 归责困境与出路

李贤森

(北京理工大学 法学院, 北京 100024)

摘要:数字政府建设中,多轮知识蒸馏技术被广泛应用于大模型的私有化部署。与传统算法相比,该技术特有的“教师模型—中间学生模型—最终学生模型”多级嵌套结构,不仅容易放大历史数据偏见,更使因果证据链在模型代际传递中发生断裂,引发行政主体确认困难、法定举证责任落空及行政裁量权受损等归责困境。面对复杂链式模型,现行用途导向与单一归责的监管逻辑面临失灵。为弥合技术与法律的错位,规制路径不应逾越法治底线干预前端市场,而应立足数字政务的公法属性,将监管重心转向内部治理与采购合规。应确立政务采购中的模型链条强制披露机制,并辅以基于合同的责任联动追偿体系,从而精准锚定过错节点。这既能保障行政主体实现有效追偿,也为人工智能行政治理提供了合法且可行的制度框架。

关键词:多轮知识蒸馏;大模型应用;行政法归责;政务采购强制披露;技术黑箱;责任联动机制

中图分类号:D922.16 **文献标志码:**A **文章编号:**2096-028X(2026)02-0117-12

随着生成式人工智能技术的迅猛迭代,大模型的架构日益复杂化且应用边界不断拓宽。多轮知识蒸馏是兼顾模型性能与算力成本的核心技术路径,其技术原理源于计算机科学领域的知识蒸馏框架,即通过教师模型向中间学生模型(简称中间模型)再向最终学生模型(简称最终模型)传递概率分布形式的“软标签”,实现大型模型向轻量化模型的知识迁移。^①与单次蒸馏不同,多轮蒸馏在教师与最终学生之间引入一个或多个中间学生,形成逐级压缩的链式知识传递结构,^②并已全面渗透至大型模型的训练与部署实践中。然而,传统行政法归责体系在应对人工智能系统时本就面临主体认定与因果关系模糊的固有困境。多轮知识蒸馏技术虽赋予了模型轻量化与低成本的优势,促使其在政府治理中被广泛采纳,但其特有的“教师模型—中间模型—最终模型”的多级嵌套与知识传递链条,进一步拉长并稀释了本就脆弱的归责链条。这种技术演进使行政行为的因果关系追溯面临前所未有的技术屏障与制度断层。

多轮知识蒸馏通过“教师模型—中间模型—最终模型”的链式路径,将庞杂的知识压缩到轻量化系统内。以政务服务智能问答系统为例,其底层往往依赖此类多轮蒸馏模型以降低算力部署成本。然而,这种链式结构具有强烈的路径依赖与隐蔽性,前序模型中的微小数据偏见或算法缺陷,会在传递中被继承甚至放大。当行政机关将其嵌入自动化决策、风险评估等公共服务场景时,这种叠加的技术黑箱直接冲击了依法行政所要求的透明、正当与可释明原则。

中国高度重视人工智能技术在公共治理领域的深度应用,明确提出将智能技术融入社会治理体系,通过部署政务辅助与决策支持系统,全面提升公共服务效能与治理现代化水平。^③将多轮知识蒸馏引发的复杂归责问题置于行政法治框架下审视,实质上是以底层算法规制赋能国家治理体系与治理能力现代化的重要

基金项目:2024年度司法部法治建设与法学理论研究部级科研项目“我国应对欧美数据法域外适用扩张的对策研究”(24SFB4023),2025年度教育部人文社会科学研究青年基金项目“欧美数据法域外管辖权冲突与中国‘合理联系原则’应对机制研究”(25YJC820020)

作者简介:李贤森,男,法学博士,北京理工大学法学院长聘副教授、研究员、博士生导师。

^① See Geoffrey Hinton, Oriol Vinyals & Jeff Dean, *Distilling the Knowledge in a Neural Network*, Computer Science, Vol.14:1, p.1(2015).

^② See Seyed Iman Mirzadeh & Mehrdad Farajtabar, et al., *Improved Knowledge Distillation via Teacher Assistant*, Proceedings of the AAAI Conference on Artificial Intelligence, Vol.34:5191, p.5191(2020).

^③ 参见陈家喜:《运用人工智能提升社会治理效能》,载《红旗文稿》2025年第13期,第19页。

命题。这一视角的引入标志着人工智能治理逻辑的结构性转向,即突破传统私法偏重事后损害赔偿与个体权利救济的局限,迈向以公共利益为导向、以公权力规范运行为核心的公法治理模式,进而强化对技术风险的前置性审查与系统性预防。

行政法视野下的人工智能治理超越了传统的事后追责逻辑,更加凸显事前风险预防、事中过程监管及行政机关在重塑治理规则中的主导地位。相较于私法归责聚焦于个体间的权利救济与损害填平,行政法归责的核心在于规范公权力运行、维护公共秩序与构筑公民基本权利体系的保障体系。因此,将人工智能的责任治理从私法救济转向公法视域,要求深刻剖析技术演进带来的结构性挑战,重构行政机关行使公权力应对技术风险的制度框架,从而确保智能技术在公共领域的应用始终被约束于依法行政的轨道之内。

一、多轮知识蒸馏技术在数字政府中的运作实践与规制挑战

(一) 政务大模型私有化部署的实证运作机制

随着数字政府建设的深入,大语言模型在政务服务、辅助决策与城市治理中的应用日益广泛。然而,千亿级参数的通用大模型在私有化部署时面临极高的算力成本与数据安全壁垒。在此背景下,多轮知识蒸馏作为一种模型压缩与轻量化技术,成为数字政务采购与部署的重要需求。该技术通过教师模型至中间模型再至最终模型的链式路径,将庞大复杂的知识压缩到轻量化的学生模型中。这一过程在大幅降低算力负担的同时,有效维持了特定政务场景下的文本理解与生成能力。

进一步而言,多轮知识蒸馏技术在数字建设中的大规模展开,是由政务数据的高度敏感性与基层行政资源稀缺性之间的结构性矛盾决定的。一方面,根据《中华人民共和国数据安全法》(简称《数据安全法》)第21条,中国对政务数据安全实施严格的分类分级保护制度,通常要求核心政务数据在安全可控环境中处理,并在实践中强调原始数据不出域、重要业务系统与公共互联网相隔离。这意味着行政机关无法直接调用部署在商业云端的大型通用模型应用程序编程接口,否则将面临数据泄露与国家安全风险。另一方面,若要在地方政府或基层执法部门的内网环境中进行大型模型的私有化本地部署,需配置算力集群,负担高昂的硬件维护成本。在这一现实张力下,多轮知识蒸馏成为了数字政务落地的重要方式。将庞大、通用的教师模型中的知识表征进行针对法律法规、地方政策、历史工单等特定政务语料的多轮次蒸馏与裁剪,最终模型的参数量可能被压缩至原始模型的百分之一甚至更低。这使得模型能够顺利部署在单台普通服务器乃至边缘计算设备上,从而契合了政务系统对数据物理隔离与低成本敏捷部署的双重需求。然而,此种以牺牲部分泛化能力为代价的深度参数压缩成为后续行政决策偏见与逻辑黑箱产生的技术根源。

在目前的行政管理实践中,轻量化人工智能模型已深入具有法律效力的决策环节。以地方政务热线智能辅助系统为例,为实现本地化的高效部署,地方政府往往采购基于知识蒸馏技术生成的轻量化政务专用模型。该模型在实践中承担着智能派单、诉求定性与公文处理等关键功能。当市民通过热线或政务应用程序提交诉求时,模型会基于自然语言理解自动判断管辖层级,并将工单派发至特定的基层执法部门。针对部分常规咨询,模型甚至能自动生成具有政府背书性质的答复意见。从公共行政到自动化行政,形式上从人类决策转变为算法决策,本质上依然是行政权的运作。^① 在此类运作中,轻量化模型不再仅仅是办公辅助工具,而是实质性地嵌入了行政调查与指令下达等具有外部行政法意义的权力运行链条之中。

各地政府采购实践中,政务大模型及相关智能系统的采购规模呈快速增长态势。深圳市政务服务和数据管理局采购的数字政府人工智能公共支撑平台项目中标金额达2306万元,旨在构建“1门户+1中心+1中台+1体系”的人工智能应用支撑架构。^② 上海市信访办公室采购的12345市民服务热线数据分析及大模型技术支撑服务项目预算金额达1128万元,明确将大模型应用赋能服务热线作为建设目标。^③ 烟台市大数据局政务人工智能大模型服务项目中标金额达289万元,涵盖大模型管理支撑工具、训练调优工具、智能体

^① 参见陈悦:《论自动化行政中算法决策风险的“人在回路”治理模式》,载《行政法学研究》2024年第4期,第39页。

^② 参见《深圳市政务服务和数据管理局数字政府人工智能公共支撑平台项目中标(成交)结果公告》,载深圳政府采购智慧(监管)平台2025年3月10日, <http://zfcg.szggzy.com/8081/gsgg/002001/002001004/002001004001/20250311/5ea54613-8ac4-4665-9edb-e457b6de1f8f.html>。

^③ 参见《上海12345市民服务热线数据分析及大模型技术支撑服务项目中标(成交)结果公告》,载中国政府采购网2025年9月25日, https://www.ccgp.gov.cn/cggg/dlfgg/zbgg/202509/t20250925_25414573.htm。

开发工具等五大支撑工具与四大服务。^① 这些采购实践表明,大模型技术已从实验性应用转向规模化政务部署,相应的归责制度需求亦日益紧迫。

(二) 多级蒸馏链条引发的行政归责实证困境

由多轮知识蒸馏支撑的运作实践虽然极大提升了行政效能,但也在行政归责领域诱发了系统性风险。在上述智能派单系统的实证运行中,知识传递过程中的有损压缩与逻辑偏移易使轻量化模型因语料识别偏差而出现系统性误判,进而导致行政机关无法及时履行法定职责,引发行政不作为风险。

近年来已出现人工智能技术影响行政决策引发行政复议的实例。例如,2025年司法部发布的行政复议典型案例中,“王某不服浙江省某市市场监督管理局不予受理撤销冒名登记行政复议案”即涉及人工智能换脸技术破解政务平台实名核验系统的归责争议。^② 复议机关认定,被申请人仅凭后台文字记载即认定申请人完成真实身份核验,未排除人工智能技术伪造可能性,属于事实不清、证据不足,最终撤销不予受理决定并责令重新审查。该案例表明,行政机关在借助智能系统作出决定时,若未能审慎审查系统输出的可靠性,就可能需承担相应的法律责任。与此同时,在行政处罚辅助决策领域,甘肃省酒泉市中级人民法院在“嘉峪关浩峰能源有限责任公司诉嘉峪关市环境保护局环保行政处罚案”中查明,环保部门依据行政处罚自由裁量辅助决策系统作出的处罚结果超过法定裁量范围,最终判决撤销该行政处罚决定。^③ 此类案例揭示了自动化决策系统输出结果可能超越法定权限的法治风险。

相对人对此类数字化行政行为提起行政复议或诉讼时,传统归责逻辑将面临适用困境。多轮知识蒸馏形成了一个极度复杂的技术供应链,在政务内网上运行的最终的轻量化模型,其决策逻辑不仅受制于本地运维平台的微调,更深层继承了上游中间模型的压缩算法及原始教师模型在预训练阶段的数据特征。在实践中,由于前序模型中的微小数据偏见或算法缺陷会在蒸馏传递中被隐蔽地继承甚至放大,行政机关在应对追责时,无法解释究竟是哪一轮的知识传递导致了派单算法的失灵。这种链式结构具有强烈的路径依赖与黑箱叠加属性。这不仅使行政行为丧失了法治所要求的可释明性,^④也直接阻断了行政机关向技术供应商进行事后追偿的事实证据链条。

二、多轮蒸馏大模型对传统行政法归责原理的冲击

从前文所述的技术原理出发,与数字行政中传统的基于逻辑规则的专家系统乃至常规的单轮生成式人工智能相比,多轮知识蒸馏技术的本质特征可概括为知识的跨代际有损压缩,其对行政法归责原理的冲击具有显著的代际差异与结构复杂性。传统自动化行政的归责难点主要在于代码翻译错误,其因果链条依然是线性的,责任主体往往锁定于单一的系统开发商。即便是普通的单轮大模型,其算法黑箱也多局限于单一主体的数据训练边界内。然而,多轮知识蒸馏打破了单一开发者对应单一模型的映射关系,形成了多主体嵌套与责任碎片化的技术供应链。上游教师模型的预训练偏见与事实错误,会通过软标签以隐蔽的方式传导给下游的学生模型,并在特定政务场景中被放大,使得损害结果与致害行为之间隔着数代算法的转换,导致传统的行政法归责要素在面对跨代际的技术链条时,出现了系统性的适用失灵问题。

(一) 链式结构解构单一行政主体的归责基础

多轮知识蒸馏模型形成的复杂链式结构,动摇了传统归责体系的重要基础——可识别的行为主体与清晰的因果链条。在行政法领域,这一挑战尤为突出,导致了主体缺位与意志空洞。当行政机关使用多轮蒸馏人工智能模型进行行政决策或提供公共服务时,如何认定行政主体成为核心问题:主体是行政机关自身,还是人工智能模型,抑或是二者协同?人工智能模型的行为是否可被视为行政机关的行政行为或其组成部分?

^① 参见《烟台市大数据局烟台市政务 AI 大模型服务项目公开招标公告》,载烟台市公共资源交易网 2026 年 2 月 13 日, <https://ggzyjy.yantai.gov.cn/jyxx/003002/003002002/20260213/ed275e5b-5640-4f3d-881d-308259da0af5.html>。

^② 参见《2025 年度行政复议典型案例》,载中华人民共和国司法部网站 2026 年 2 月 12 日, https://www.moj.gov.cn/pub/sfbgw/fzgz/fzgzxzzf/fzgzfygz/202602/t20260212_531732.html。

^③ 参见“嘉峪关浩峰能源有限责任公司诉嘉峪关市环境保护局环境保护行政处罚案”,甘肃省酒泉市中级人民法院(2018)甘 09 行初 2 号行政判决书。

^④ 参见王文玉、杨丽芳:《自动化行政的法治风险样态与治理路径》,载《合肥工业大学学报(社会科学版)》2025 年第 5 期,第 138 页。

这些问题都要求对行政行为概念进行重新理解和范围拓展。

与传统自动化行政系统或单轮生成式人工智能相比,多轮知识蒸馏在主体认定上的根本区别在于,其打破了单一开发者对应单一模型的线性映射关系,导致行政意志剥离与归责主体碎片化的困境。^①在单轮大模型应用中,归责主体通常局限于系统开发商与行政机关;但在多轮蒸馏架构下,基础通用大模型的提供者仅输出原始知识表征,中间模型的开发者进行领域化剪枝与微调,^②最终的轻量化政务模型才直接作出行政输出。跨代际的多主体协同意味着,任何一项具体的自动化行政决定,都不再是单一主体意志的完整体现,^③而是历经多代算法主体裁剪与重构后的产物。多轮知识蒸馏的链式演化使得行政法上权责一致的单一主体预设陷入失灵。当行政损害发生时,由于行政意志在多级蒸馏中被逐渐稀释,传统法律语境下基于行政行为及其产生损害的对应性、可归因性判断丧失了事实基础,^④导致归责主体难以确定。

这种主体认定困境直接延伸至行政责任承担机制的挑战与调适。当多轮蒸馏人工智能模型在行政管理或公共服务中造成损害时,行政机关是否应承担行政赔偿责任、行政补偿责任,相关公务员是否应承担行政纪律责任,都成为亟待明确的问题。这涉及行政机关的过错认定、控制力边界及责任分担机制,尤其是在公私合作模式下,如政府向人工智能提供方购买人工智能服务,二者的责任划分更为复杂。多轮知识蒸馏形成的模型生成链条,在技术维度呈现深度依赖与功能嵌套特征的同时,亦在法律层面引发责任认定的去中心化与多层稀释效应,在这种情况下政府通过法律进行直接规制的手段将受到制约。^⑤

相较于传统单一模型、单一开发者场景,多轮蒸馏大模型的责任分配机制更为复杂,其法律规制面临因果链拉伸、行为主体扩展、责任边界模糊等问题,导致责任主体难以确定、归责路径难以厘清。在实践中,这种多元参与者角色的模糊更是引发了主体确认的难题。行政主体和行为归责的模糊化,不仅是程序上的障碍,更是对行政法中行政主体这一核心概念的挑战,若谁在行动和什么构成行动变得不确定,那么建立在公共权力行使基础上的整个行政法体系将面临基础动摇的风险。这预示着行政法可能需要发展出新的分布式主体或集体责任理论,^⑥以适应人机协同和多利益攸关方参与的复杂现实。

(二) 技术黑箱导致法定举证责任与因果链条断裂

多轮知识蒸馏将模型训练的演化进程从传统的一次性生成拓展为多层级、多阶段的知识迁移链式路径。这导致传统因果归责逻辑在复杂模型链条中的适用失灵问题。行政法中的因果关系认定,在人工智能模型复杂链条下,面临事实不能或法律不能的挑战。例如,要解释一个具有成千上万个决策树的危害风险评估工具算法的特定决策,在技术上是切实际的。^⑦在传统的数字行政或单轮大模型引发的争议中,行政机关尚且可以通过调取原始训练语料或单一模型的决策树来完成初步的举证义务。然而,多轮知识蒸馏的本质是知识的有损压缩与跨代际传递,最终的轻量化模型仅继承了前序模型的软标签即概率分布特征,而非原始训练数据。这意味着前序模型中的数据偏见或算法过错在向后传递时,已经被剥离原始的数据载体。当相对人对自动化行政决策提起复议或诉讼时,作为被告的行政机关即使完全公开最终政务模型的参数,也无法逆向还原出上游基础模型的原始致害机理。多轮知识蒸馏模型的代际传递造成的证据链断裂,使得行政诉讼法规定的由行政机关承担举证责任的制度面临无证可举的困境,从根本上动摇了行政法事实认定与法律关系构建的基础。

行政法要求行政行为公开透明,以保障公民知情权、参与权和监督权。因此,政府所用人工智能模型的

① 参见叶昶秀:《政府算法决策的归责困境与治理路径》,载《华中科技大学学报(社会科学版)》2026年第2期,第138页。

② “剪枝”是计算机科学中的规范概念,旨在通过减少模型中的冗余参数和连接实现模型的压缩和加速。参见王琳、宋权润等:《神经网络滤波器剪枝技术研究综述》,载《计算机工程与应用》2026年第2期,第1页。

③ 生成式人工智能将不可察觉地抢占人的主体性,本质上也在高度影响主体意志。参见张爱军、乔运涛:《从媒介工具性到技术本体论:西方生成式智能的数字帝国主义风险》,载《未来传播》2024年第1期,第34页。

④ 参见汪青松:《大模型运行损害的归因范式与责任机制》,载《东方法学》2025年第4期,第132页。

⑤ 参见周文康:《生成式人工智能安全治理的反身法路向——以DeepSeek为例》,载《新疆社会科学》2025年第2期,第74页。

⑥ 参见牟春花:《算法物化、集体法律责任与分布式法律责任:规制算法的三重递进》,载上海市法学会编:《上海法学研究》第7卷,上海人民出版社2022年版,第213页。

⑦ 危害风险评估工具通常指代一种用于评估和预测潜在危害和风险的系统或方法。See Marion Oswald & Jamie Grace, et al., *Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and ‘Experimental’ Proportionality*, Information & Communications Technology Law, Vol.27:223, p.231 (2018).

算法透明度,^①包括模型创建方式、训练数据、决策逻辑等,应基于算法可公开性与可解释性满足行政公开原则,提升行政行为公信力。^② 算法透明度有助于打开人工智能黑箱,以更好地了解模型如何作出决策。然而,多轮蒸馏模型的决策过程具有高度的隐蔽性与非线性特征,其输出内容是多代模型参数更迭与协同演化的综合结果。生成路径的复杂性导致模型具体行为难以直观还原,直接阻碍传统法律体系中由行为到后果的证据链构建。在传统侵权与产品责任学说框架下,原告需就加害行为、损害结果及因果关系承担举证责任。^③ 但在多轮蒸馏场景中,模型的学习参数、训练数据、蒸馏逻辑与输出内容之间缺乏直接可追溯的因果路径。即便技术上可定位某输出内容源自特定训练数据片段,该数据是否构成模型产生特定输出内容的关键致因仍难以证实。这种不可归因性严重阻碍模型行为责任的认定,导致传统举证规则陷入失灵。

行政法本身在某些领域就存在黑箱问题,如行政裁量权的行使过程不透明,进而导致行政裁量的滥用。^④ 当行政机关引入人工智能进行决策时,可能导致行政黑箱与人工智能黑箱的叠加,进一步削弱行政行为的可解释性、可预测性和可问责性。多重嵌套的技术黑箱将直接危及行政诉讼中的举证责任分配机制。《中华人民共和国行政诉讼法》确立了由被告(行政机关)对其所作的具体行政行为合法性负举证责任的原则。在数字化行政争议中,当相对人对自动化决策结果的合法性提出质疑时,行政机关必须向法院充分释明其所作决策的事实根据与算法逻辑。然而,在多轮蒸馏的复杂链条下,作为使用者的行政机关往往只掌握最终模型的输出结果,无法解释上游教师模型及中间蒸馏环节的决策机理。算法的不可解释性使得行政机关在数字化争议中面临极大的举证不能风险,极易因未尽法定说明义务而承担败诉后果。多轮知识蒸馏的黑箱困境不仅阻碍了表层的技术审查,更从根本上动摇了行政法事实认定与法律关系构建的证据基础。在行政决策内在机制不透明的情况下,法律审查和问责的基础将被弱化。

司法实践表明,法院在审查自动化行政决策时往往面临技术性证据的认定困难。以“上海市何凯诉黄浦交警支队案”为例,^⑤当事人从声学专业角度对电子监控设备的波段曲线及标记准确性提出质疑,但法院在审理中明确表示:“对于技术问题,行政审判的审查强度有限,且专业问题会因为观点角度的不同存在不同的理解和争议,司法审查不宜过多干预。”法院对于声呐设备的声源定位原理、技术误差范围等实体性问题未予实质审查。这一审查路径反映了一个结构性问题:依据《中华人民共和国行政处罚法》第41条,司法机关对电子技术监控设备的审查主要围绕“设置目的是否正当、标志是否明显、程序是否合法”展开,审查维度以“技术性正当程序”为边界,未延伸至算法规则与裁量逻辑的内在合理性。学界将此现象概括为技术壁垒与算法黑箱的双重困境。^⑥ 自动化量罚系统的算法规则、算法决策理由与自动化决策过程构成司法审查的对象,但法院因专业能力与制度工具的限制,仅对“形成过程”“明显性”“实际效果”进行了审查。^⑦ 司法机关在面对技术黑箱时,存在一定程度的审查回避倾向,在执法机关借助技术系统生成处罚结果的情况下,法院往往只能在程序合法性层面进行审查,而难以对技术决策的实体合理性作有效判断。自动化决策系统的算法逻辑难以进入司法实质审查视野,相对人即使胜诉亦无法解决算法错误的责任归属问题。

(三) 算法依赖影响行政裁量权的正当性

人工智能模型对行政权力的深刻冲击,还体现在对行政裁量权正当性的影响。在行政裁量领域,法律赋予行政机关基于具体情境进行利益衡量与价值判断的权力,行政裁量因而具备实现个案正义的法律品格。^⑧

^① 人工智能透明度原则不能只是智能系统体现自身意愿的单向度的任意信息披露,换句话说,透明度原则具有信息披露的形式透明要求,更具有面对不同受众可解释性需求的实质规范内容。参见张永忠:《论人工智能透明度原则的法治化实现》,载《政法论丛》2024年第2期,第126页。

^② 参见于帆:《自动化行政中算法的可公开性及其范围》,载《行政法学研究》2024年第1期,第162页。

^③ 参见杨立新:《产品责任损害赔偿范围问题研究——〈侵权责任编司法解释(一)〉第19条的理解与适用》,载《中国应用法学》2024年第5期,第9页。

^④ 参见卢政峰:《论行政自由裁量的“解释性控权”》,载《法律方法》2021年第3期,第266页。

^⑤ 参见“何凯诉上海市公安局黄浦分局交通警察支队交通行政处罚案”,上海市高级人民法院(2019)沪行终204号行政判决书。

^⑥ 参见于洋:《自动化行政处罚的司法审查——以道路交通领域为例》,载《比较法研究》2024年第4期,第197页。

^⑦ “形成过程”审查,即审查该套设备系统投入运行前是否经过了充分的论证、检测;“明显性”审查,即审查该套设备系统是否存在明显的违反逻辑和科学性的情形,而行政机关对此难以解释;“实际效果”审查,即审查该套设备系统投入运行后,是否存在明显大量的异议(复议或诉讼),导致产生科学性上的合理怀疑。参见“何凯诉上海市公安局黄浦分局交通警察支队交通行政处罚案”,上海市高级人民法院(2019)沪行终204号行政判决书。

^⑧ 参见赵龙:《自动化行政裁量的包容性规制》,载《行政法学研究》2025年第1期,第57页。

但是经过多轮蒸馏的政务大模型被引入行政处罚幅度建议、社会救助资格实质审查等裁量决策场景时,行政机关也易出现自动化偏见与裁量怠惰的问题,不利于实现个案正义。

与一般人工智能引发的自动化偏见不同,多轮蒸馏大模型对行政裁量权正当性的影响,源于其底层算法架构对特殊情境的系统性忽视。行政裁量权的核心在于通过利益衡量实现个案正义,尤其强调对边缘及特殊行政情节的考量。单轮大模型尚有可能保留相对全面的数据特征,但在多轮知识蒸馏的过程中,轻量化模型为追求运行效率,其算法逻辑天然倾向于拟合大模型中的主流普遍特征,而不可避免地剔除或弱化处于边缘概率的复杂情境。这就导致蒸馏后的政务模型在面对具有特殊情节的行政相对人时,极易输出高度同质化且一刀切的决策建议,并往往带有隐性的偏见与歧视,干扰行政人员最终决策。^①这不仅诱发了行政人员的裁量怠惰,更在算法压缩设计之初就挤压了行政法追求个案实质公平的技术空间。当一线执法人员长期依赖此类输出结果时,法定的人工复核往往流于形式,最终导致行政机关的实质裁量权被隐藏在代码深处的压缩规则所影响,严重动摇了行政权行使的合法性与合理性基础。这要求行政机关在运用人工智能时,必须保留实质性的“人工干预权”和“最终解释权”,确保算法仅作为事实查明的辅助工具,而不能让渡行政行为的最终价值判断权。^②

多轮蒸馏技术的“技术中立性”概念,在公共治理语境下具有误导性。尽管技术本身在设计上可能是中立的,但其应用方式和黑箱性质等内在特性却与行政法的公平、透明等基本要素存在固有冲突。这揭示了技术效率及中立性与公共法律规范要求之间存在的深层矛盾。行政法不能简单地将人工智能视为一个中立工具,而必须积极审视其内部机制及其对社会的影响。

传统归责体系在多轮蒸馏面前展现出了较大局限性。合法行政、合理行政、比例原则和信赖保护原则等行政法基本原则,是规范行政行为、保障行政法治的核心约束。行政法基本原则应是指能够统领全部行政法规范的原则,既能够调整所有行政法主体,也能够调整所有行政法行为。^③那么在人工智能背景下,行政法基本原则的适用边界是否需要重新定义?例如,人工智能决策的效率与公平之间如何体现比例原则?人工智能的不可解释性如何与行政行为的可解释性要求相协调?《中华人民共和国个人信息保护法》(简称《个人信息保护法》)中的自动化决策拒绝权在行政法领域如何适用、其重大影响的判断标准如何确定?^④这些问题是对行政法理论的深层挑战,意味着行政法需要对自身重大基础原则(如正当程序原则)进行根本性重构,以在自动化决策时代保持其规范效力。^⑤应对人工智能对行政裁量权 and 基本原则的挑战,不仅仅要确保合法性,更要维护行政法本身的规范核心。如果人工智能削弱了人工控制、公平性和可解释性,那么其将侵蚀公共行政的合法性基础。这要求决策者主动介入,塑造人工智能在公共行政中发展和应用的形态,确保技术进步与法治价值相协调。

三、现行多轮蒸馏模型责任监管的制度性缺失

(一) 用途导向的外部监管难以穿透技术供应链

当前人工智能监管体系在构建归责机制时,多以“用途导向”原则为基准,即依据模型实际应用场景与最终用途界定法律责任。^⑥该逻辑在早期人工智能模型功能单一、行为路径可控的语境下具备合理性,但在多轮知识蒸馏架构中,正逐渐显现监管滞后与规范效力不足的问题,难以应对复杂模型链条引发的责任扩散与风险隐匿现象。

“用途导向”原则的核心困境在于:其将监管视角局限于模型外部部署环节,忽视了模型生成机制中知识责任的层级传导结构。即便最终模型作为面向市场的终端载体,其训练语料、行为模式及判断逻辑仍从根

① 参见刘飞、吴辉:《智能向善:人工智能价值对齐的人文建构》,载《成都理工大学学报(社会科学版)》2025年第5期,第35页。

② 参见程关松、苗运卫:《相对人参与算法行政的困境与出路》,载刘艳红主编:《东南法学》第4辑,东南大学出版社2021年版,第46页。

③ 参见吕成:《法典化背景下行政法的双重体系建构》,载《理论月刊》2025年第6期,第150页。

④ 参见翁明杰:《智能行政中自动化决策拒绝权的证成与适用》,载《财经法学》2025年第1期,第149页。

⑤ 参见唐安然:《自动化行政中正当程序原则的重构》,载《比较法研究》2025年第5期,第186页。

⑥ 人工智能系统的输出并非完全独立于人类设计和训练过程,这意味着责任的分配可能需要综合考虑开发者、使用者及系统本身的特性。在数字时代,随着信息形式日益多样化,传统的物质形态已不再是判断其性质的唯一标准。相反,信息的功能、用途及对用户行为的实际影响可能更为关键。参见李雅男:《生成式人工智能的法律定位与侵权归责路径》,载《比较法研究》2025年第3期,第73页。

本上依赖教师模型及中间模型。在蒸馏过程中,知识的压缩、抽象与表达重构呈层级推进,学生模型往往继承甚至放大教师模型潜在的判断偏差与结构缺陷。若缺乏对教师模型环节的监管介入,仅以用途作为评价依据,则低估了整个链条中其他主体的法律义务。同时,“用途导向”掩盖了通用模型与专用模型间的法律结构差异。实践中通用模型常被蒸馏为小型化、垂直化的场景专用模型,而“教师—学生”结构可能由不同主体实施控制。若法律仅聚焦最终用途场景,将实质性削弱通用模型提供者在知识源头的规范责任。以 DeepSeek 等企业为例,其通过多轮蒸馏技术将通用模型转化为司法或医疗领域的专用模型时,未重新训练原始语料即赋予模型场景功能,致使算法偏见、知识漏洞等风险未经有效审查便被继承。

同时,以用途为中心的监管可能诱发法律规避行为。以开发者为例,部分开发者刻意维持开发脱钩表象,仅对外提供蒸馏模型的压缩版本,宣称不介入下游部署以规避行政责任;同时,若按照产品责任界定及其风险抗辩规则,开发者极易规避侵权责任,人工智能使用者的利益也难以得到保护。^① 这种现象可能会在技术与商业层面消解监管约束效力,形成监管盲区,也不利于促使服务提供者履行注意义务并主动规避监管风险。此外,当前人工智能系统的高度开放性与演化性,使得模型用途在实际运行中存在动态变异可能。一款初始设计为文本摘要的模型,可能通过与用户的持续交互,习得性地演化为问答系统或推荐工具,导致实际用途与开发初衷背离。在此情形下,基于初始用途设定的归责路径,将无法覆盖人工智能模型全生命周期的法律风险。^②

对此,行政法应重构行政许可与备案制度,拣选出可适用于人工智能行政管理领域且无法从行政法基本原理中推导出的制度。^③ 应引入全生命周期行政许可或备案制度,将监管重心前移至模型开发、训练、蒸馏等早期阶段。对高风险人工智能模型,特别是作为基础的教师模型,应要求其在投入蒸馏链条前进行行政许可或备案,并定期更新,确保其训练数据、算法逻辑、潜在偏见等符合行政法要求。此外,欧盟《人工智能法案》确立了以风险为基础的四级分类监管模式,将人工智能系统划分为不可接受风险、高风险、有限风险和最小风险四个层级,对不同层级施加了差异化的合规义务。^④ 该框架对于多轮蒸馏模型在政务场景中的规制具有重要参考价值。一方面,其风险分级理念可为行政机关提供差异化监管的规范参照——对涉及公共安全、重大福利分配的高风险自动化决策系统,应施加更严格的审查评估义务;另一方面,其确立的基本权利影响评估制度要求部署者在部署高风险人工智能系统前评估此举对基本权利的影响,^⑤ 这与中国政务采购中建立算法影响评估制度的需求相契合。然而,该框架亦存在明显局限:其风险分类以静态清单为主,难以适应多轮蒸馏模型在不同政务场景中风险水平的动态变化;合规成本较高,可能加重基层行政机关的负担;其基本权利保护路径主要依托风险预防逻辑,与基本权利保护的合法与非法二元判断逻辑之间存在协调问题。因此,在借鉴欧盟经验时,应结合中国行政体制特点,采取场景化动态风险评估路径,根据人工智能系统在行政决策中的具体角色确定风险级别,在法律移植过程中重视制度背景的差异。

(二) 单点问责模式与链式协同开发结构不匹配

多轮知识蒸馏在模型演化进程中构建起动态生成链条,其显著特征为生成主体的非单一性、生成过程的连续性与生成结果的叠加性。但是,中国现行法律体系中关于责任主体的认定规则,多以单一行为体与一次性行为为预设前提,难以有效涵盖链式模型生成所形成的复合责任情况。这种规范预设与技术特性的脱节,不仅导致法律规制对象与技术生成链条的结构性错位,更影响责任分配的合理性,削弱法律规则的执行效能与可预测性。

行政监管通常以产品归属或服务提供者作为管辖基础,但蒸馏链条生成的模型可能由多个开发者、平台甚至开源社区协同完成。若仅将最终模型输出者作为监管对象,将导致链条关键节点陷入监管盲区,削弱整

① 参见窦海阳:《人工智能产品侵权归责何以脱离产品责任?》,载《中外法学》2025年第2期,第330页。

② 参见李贤森、李坤锦:《全生命周期视野下人工智能数据出境安全监管的具体场景、重点风险及体系构建》,载《河北法学》2026年第2期,第101页。

③ 参见陈鹏:《从本质主义到原旨主义:行政许可界定路径之反思》,载《法律科学》2025年第5期,第144页。

④ See Regulation (EU) 2024/1689 of the European Parliament and of the Council, euorgip, https://euorgip.fr/wp-content/uploads/2024/12/Regulation_harmonised_rules_AI_2024.pdf.

⑤ 参见李贤森、李庭凯:《论人工智能系统法律风险规制——以欧盟〈人工智能法〉为例》,载《北京科技大学学报(社会科学版)》2025年第5期,第87页。

个监管体系的严密性。在类型化视角下,单一化的侵权归结方式不能周延地约束技术生成链条的多元演进路径,^①若不能及时构建识别链式生成特征、容纳多元责任分布的新型法律框架,将加剧责任归属混乱,弱化法律对人工智能生成内容的实质约束力。以侵权责任为例,传统责任主体需具备明确的可识别性与行为自主性,但人工智能的行为仅具有有限的自主性和可解释性,且模型蒸馏过程中的系统联系加剧了决策主体的复杂性,这给侵权责任界定带来了全面挑战。^②

在多层蒸馏语境中,最终模型作为输出载体,其生成行为深度依赖前级模型传递的知识表征,而这种知识传递过程难以被界定为直接控制或委托代理关系。传统侵权认定强调从行为到后果之间的因果关系,但蒸馏链条呈现非线性、多节点扩散特征,部分中间模型虽未直接参与最终内容生成,却可能因知识偏差或风险传导实质影响结果,这一现象在最新的“掩蔽生成蒸馏”模型(Masked Generative Distillation,简称MGD)中更为突出。^③它打破了传统蒸馏中学生模型直接模仿教师模型特征的线性映射逻辑。MGD 通过对学生模型的特征进行随机掩蔽,强制其利用残缺特征去反向生成和还原教师的完整特征,虽然在计算机科学意义上提升了模型的表征能力与鲁棒性,但在法学视域下,其高度非线性的知识重构过程进一步加深了算法的黑箱属性,使得原始数据偏见在随机掩蔽环节的扩散路径变得更加隐蔽和不可预测。鉴于此,法律层面的行为控制力认定标准难以精准适配模型层级复杂化引发的责任交叉问题。

传统行政法体系强调根据具体任务在组织间独立的管辖分配来划定彼此的责任范畴,缺少对行政权数字化集成行使下复合责任要求的回应。^④行政法应突破传统行政法中单一行政主体的观念,探索行政主体多元化认定。在人工智能辅助行政决策中,行政机关、人工智能模型开发者、平台运营者等多元主体可探讨其在行政法上的共同行政责任或连带行政责任。在传统的单向度行政管理中,责任主体相对明确。但在数字政务采购与应用场景中,行政机关引入多层蒸馏模型辅助决策,一旦该决策侵犯相对人合法权益,行政机关必须首先对外承担国家赔偿或行政补偿责任。在完成外部责任的承担后,行政主体需依据政府采购合同或产品责任法,向提供技术服务的多元主体进行内部追偿。然而,面对多层蒸馏的复杂技术供应链,行政机关难以准确定位损害结果的具体成因源于基础教师模型的语料污染、中间蒸馏环节的算法过失抑或终端平台的运维不当。此因果链条的断裂,导致行政机关在启动内部追偿时缺乏确凿的证据支撑与法定追偿依据,实质上被动承担了技术链条运转的所有潜在风险。

(三) 数字政务公私合作中的追偿与约束机制滞后

多层知识蒸馏的链式架构不仅加剧了法律归责的技术难度,更在实践中形成了广泛的责任规避空间。开发者可借助中间模型的“传递者”角色,刻意规避最终输出结果的法律风险。由于法律尚未建立清晰的责任分配规则,模型演化链条中的责任主体识别存在显著的主观性与裁量空间。这种技术与制度的错位,为部分参与主体规避审查义务与法律风险提供了现实可能。

在实际应用场景中,诸多模型发布者倾向于通过开源渠道发布多层蒸馏后的最终模型,以非商业使用或社区贡献名义,弱化对模型潜在输出结果的直接控制责任。随着技术供应链的精细化发展,“碎片化责任”策略日益普遍——教师模型供应者、数据整理者、中间模型优化者与终端部署者分属不同主体,缺乏统一的合规义务约束机制。这一过程中,原本应在链条中设立的责任定位点被虚置,法律可及性随之削弱。当前规范体系的回应明显滞后,《中华人民共和国民法典》《中华人民共和国网络安全法》《个人信息保护法》《数据安全法》等法律,以及《新一代人工智能伦理规范》《生成式人工智能服务管理暂行办法》等专项规制文件,均未对多层蒸馏的责任划分作细致规定。现行规范大多仍围绕“模型训练方”或“服务提供者”在单一主体维度进行设计,对链式模型生成的复杂结构特征缺乏回应能力。制度更新滞后与技术迭代加速的矛盾,不仅使

① 参见刘成杰:《类型化视角下人工智能侵权责任的归结思路》,载《法律适用》2025年第3期,第117页。

② 参见陈晨:《人工智能侵权中的责任主体及归责原则》,载《中外法学》2025年第2期,第366页。

③ 与以往的特征蒸馏方法试图让学生尽可能模仿教师的特性不同,MGD认为并不需要直接模仿教师就可以提高学生特征的表达能力。MGD首先对学生的特征进行随机掩蔽,然后通过一个简单的块,使用被掩蔽的特征生成教师的完整特征。由于在每次迭代中都使用随机像素,所以在整个训练过程中所有像素都会被使用,这意味着特征会更加鲁棒,即表示系统的抗干扰能力也会得到提升。See Yang Zhendong & Li Zhe, et al., *Masked Generative Distillation*, in Shai Avidan & Gabriel Brostow, ed al. eds., *Computer Vision—ECCV 2022*, Springer, 2022, p.53-54.

④ 参见展鹏贺:《数字政府协同原则的规范内涵及制度展开——基于行政组织关系的视角》,载《行政法学研究》2025年第5期,第90页。

治理手段难以应对风险转移与责任分散的现实问题,更弱化了法律对模型开发行为的预防性约束。^①

当前数字政务建设在制度设计上陷入了公权力技术转移的合规盲区,尚未建立与公权力向算法模型的实质让渡相匹配的监管与问责机制。传统的政府采购法体系与商业合同范本多指向实体货物采购或常规软件开发设计,其服务等级协议往往只关注系统的运行稳定性与响应指标,而对模型蒸馏过程中可能产生的语料污染、知识幻觉或算法歧视缺乏前瞻性的违约责任界定。实践中,各地政务人工智能采购项目的合同条款亦暴露出这一制度短板。某市数据局在 DeepSeek 大语言模型本地化部署采购中,曾因“符合专业条件的供应商不足三家”而废标,侧面反映出该技术领域供应商准入标准可能尚未被市场广泛接纳;^②而深圳市数字政府人工智能公共支撑平台等大型采购项目虽投入规模庞大,但其合同条款对模型蒸馏过程中可能产生的语料污染、知识幻觉或算法歧视同样缺乏针对性的违约责任界定。这一现象具有普遍性:当前政务人工智能采购合同多沿用传统软件采购模板,未能针对多轮蒸馏模型的技术特性设计专门的质量保证与责任分配条款,导致行政机关在面临算法致害时难以依据采购合同进行有效追偿。在公私合作或政府购买服务模式,提供多轮蒸馏模型的技术企业往往以商业现货软件的免责条款或算法商业秘密为由,拒绝承担因模型输出错误而引发的公共管理责任。一旦人工智能辅助生成的行政处罚决定或行政许可被司法机关以事实不清或程序违法为由撤销,行政机关在对外承担国家赔偿法项下的赔偿责任后,往往发现在现有的采购合同中难以找到针对算法致害的直接追偿条款。公法责任与私法合同约束严重错位,不仅导致公权力实质上在为技术提供方的算法过错承担最终风险,也使得技术供应商缺乏优化蒸馏链条与剔除底层偏见的内在动力,最终形成技术红利归于商业主体而法律风险留存至公共管理系统的不均衡治理局面。

四、数字政务中多轮蒸馏大模型应用的行政治理路径

(一) 确立数字政务采购中的多轮蒸馏链条强制披露机制

面对多轮知识蒸馏引发的责任隐匿与追偿困境,突破口不应在于通过普遍性的行政许可干预技术市场,而应立足于数字政府建设的内部规范,确立政务人工智能采购的强制性蒸馏链条披露机制。行政机关在采购或引入生成式人工智能服务时,应将蒸馏链条透明度作为核心的准入审查标准与合同义务。具体而言,要求技术供应商在投标或服务接入前,向采购机关定向披露其模型的层级演化信息,包括但不限于基础教师模型的数据源合法性说明、多轮蒸馏的算法传递逻辑、各级学生模型的风险评估报告等。^③通过将技术黑箱的解剖权前置于采购与准入阶段,行政机关能够提前锁定链条上各环节的技术提供方,为未来可能发生的内部责任追偿与纠纷解决奠定确凿的证据基础。

在这一制度构建中,法国的相关经验可提供有益参照。法国 2016 年《数字共和国法》在全球率先将算法透明度义务嵌入行政法典,规定基于算法处理的个人决定须向当事人明确声明,当事人有权要求获知算法处理的规则及其主要特征。^④ 法国最高行政法院进一步明确,仅当算法及其逻辑能够向受影响者解释时,方可作出完全自动化决策,正因如此,完全自动化决策在日常使用中非常罕见。^⑤ 此外,法国 Etalab 工作组发布的公共部门算法问责六项原则——承认、一般解释、个人解释、理由、发布、允许争议,构成了从告知到救济的全链条问责框架。然而,该制度在实施层面存在明显不足:法律适用范围存在大量例外,且未设立中央算法登记系统,导致实际实施率较低,算法透明度在实践中很少得到真正落实。^⑥ 对中国而言,法国经验的价值在于其将算法透明度义务嵌入行政程序法的立法技术及其构建的全链条问责原则体系中,但其低实施率

① 参见李洋、薛澜:《颠覆、调适与协同:责任伦理视域下生成式人工智能的多主体治理机制研究》,载《电子政务》2025 年第 7 期,第 41 页。

② 参见《岳阳市 DeepSeek 大语言模型本地化部署政府采购服务项目(2026 年 3 月~2029 年 3 月)废标公告》,载岳阳市公共资源交易网 2026 年 3 月 11 日, https://ggzy.yueyang.gov.cn/56114/56131/56135/56257/content_2365680.html。

③ 参见杨一健、王锡锌:《数字行政的算法风险及法律规制:法国经验与中国意义》,载《行政法学研究》2024 年第 4 期,第 32 页。

④ See French Digital Republic Act, Art.47-48.

⑤ Conseil d'État, « Intelligence artificielle et action publique: construire la confiance, servir la performance », 31 mars 2022, https://conseil-etat.fr/content/download/175811/file/SRE_etudePM%2520IA_avec%2520hyperlien.pdf.

⑥ See Soizic Pénicaud, *Making Algorithm Registers Work for Meaningful Transparency*, IA Ciudadana, <https://iaciudadana.org/wp-content/uploads/2025/03/Report.pdf>.

的教训亦提示,仅有法律宣示而无制度保障难以实现预期效果。因此,中国在设计政务采购披露机制时,应建立具有强制约束力的审查程序与责任后果,避免重蹈法国“制度形式化”的覆辙。

1. 将蒸馏链条透明度设为政务采购的核心准入要件

在数字经济时代,随着算法在国家安全、公共利益、个人权利等领域影响的深入,算法治理机制将扩展并应用于更加广阔的领域。针对数字政府建设,国家应要求人工智能系统的决策过程可解释、可审查,以增强公众信任并促进公平性,^①为此应建立详细的蒸馏链条披露审查制度。该机制以模型层级信息穿透为核心,要求技术供应商在参与政务人工智能项目投标或服务接入前,向采购机关提交完整的技术链条材料,例如模型版本迭代说明、“教师—学生”模型拓扑结构、软标签生成与迁移机制、数据源构成及训练参数配置、开发主体与应用主体的角色关系图谱等。将“蒸馏链条透明度”纳入政务采购的准入范畴,尤其针对生成式政务模型需实施精细化分类与分层审查。^②对采用多轮蒸馏的复杂模型,应施加更严格的信息披露义务与“链式说明”要求,明晰每级蒸馏的输入输出信息与模型责任主体。这一制度设计可有效提升政务模型系统的可观测性,为行政机关内部的责任追踪与风险识别奠定事实依据。

2. 明确审查主体与强制披露的具体范围

首先,应明确蒸馏链条披露的审查主体。在数字政务采购场景中,审查主体不再是单一的市场监管部门,而应由具体采购或使用人工智能的行政机关,或本级政务服务与数据管理局牵头,并联合技术专家委员会进行实质性评估。披露范围应覆盖所有拟用于公共服务、自动化行政决策或具有社会管理功能的生成式人工智能模型。

其次,需细化披露材料的程序与内容。供应商的响应清单应包括但不限于:原始教师模型的技术白皮书,训练数据来源及安全性证明,各轮学生模型的性能评估报告,偏见、幻觉风险等潜在风险评估报告,蒸馏链条中开发者、提供者、部署者各方主体的内部责任分配声明。披露机制应贯穿采购的全流程,包含事前准入审查、运行期重大变更报备及年度合规报告制度,以确保披露信息的时效性。

最后,行政机关应具备对披露材料进行实质性审查和技术评估的能力,并建立差异化的分级披露与审查机制。^③考虑到技术供应商对算法源代码与模型权重主张商业秘密保护的正当诉求,政务模型的链条披露不应是无差别的全面公开。对于不涉及公民人身权与财产权的常规辅助类模型,可仅要求披露外部技术参数,即训练数据分布、测试集准确率及蒸馏策略等外部技术特征。然而,对于涉及公共安全、司法决定、社会信用评级及重大福利分配的高风险自动化决策大模型,行政机关必须引入算法影响评估制度。

在这一监管层级,应当要求技术供应商在签署保密协议的前提下,向政府组建的独立技术专家委员会或第三方算法审计机构进行深度源码级披露。审计范围应当穿透至原始教师模型的价值观对齐策略、软标签的生成日志及中间模型剪枝与量化过程中的特征丢失清单。只有在专家委员会出具未发现系统性偏见与重大合规风险的审计报告后,该模型方可获准接入政务专网进行实战部署。这种精细化的分级披露机制既尊重了参与市场竞争企业的知识产权,又有效守住了数字行政权力运行的核心安全底线。

3. 建立基于合同违约与信用档案的履约监管机制

在透明度与公开方面,通过算法的公开与解释能够实现算法的高度透明,进而达成“全景敞视主义”下的监督效果。因此,应明确披露信息的内部留存与适度公开方式。对于涉及公民基本权利的政务人工智能模型,其核心算法逻辑、风险评估结果等应向公众适度公开,增强智能化行政过程的透明度。此外,建立健全技术供应商的履约档案与信用体系至关重要。建议借助区块链等新型技术构建可信记录体系,为政务模型的迭代过程提供不可篡改的时间戳与责任绑定机制。

同时,建立政务人工智能供应商信用档案库,记录其信息披露情况、违约记录及引发的行政争议等,并与政府采购信用体系挂钩。对于违反披露义务、隐瞒教师模型缺陷、提供虚假信息或因模型问题导致行政机关对外承担赔偿责任的,行政机关可依据采购合同追究其违约责任,采取扣除保证金、解除合同、列入政府采购

^① 参见刘绍宇:《论数字政府中人工智能的法律地位与规范体系——以人工智能立法为背景》,载《行政法学研究》2025年第1期,第47页。

^② 参见唐曼:《自动化行政许可审查的法律控制》,载《现代法学》2022年第6期,第98页。

^③ 参见王文玉:《算法嵌入政府治理的优势、挑战与法律规制》,载《华中科技大学学报(社会科学版)》2021年第4期,第33页。

严重违法失信行为记录名单等惩戒措施,限制其未来的市场准入。基于数字政务内部合规的监管模式不仅要求行政机关具备更强的技术审计与跨部门协同能力,还使得原本孤立的技术审查与政府合同管理、信用监管紧密结合,形成逻辑闭环的人工智能行政治理体系。

(二) 构建基于采购合同的模型内部追偿联动机制

在多轮知识蒸馏架构中,模型行为生成机制贯穿教师模型、中间模型与最终模型的知识重构过程,行政机关在对外承担国家赔偿责任后,难以向单一的技术供应商进行精准追偿。当前,仅将最终模型部署方作为唯一违约责任主体的做法,已无法应对链式蒸馏中的风险转移难题。

对此,行政机关在构建政务人工智能采购体系时,应引入责任联动内部追偿机制。其核心逻辑在于突破合同相对性壁垒,在采购协议中明确要求最终供应商(最终模型提供方)对其上游(教师模型)的知识缺陷承担连带担保责任;或者通过多方协议,将不同层级模型的开发者共同纳入政务合规供应链。若政务大模型出现严重的歧视性输出或幻觉导致相对人权益受损,行政机关在承担外部公法责任后,可启动内部追偿。若经技术审计判定,最终模型的违规行为系由基础教师模型诱导训练且蒸馏路径存在高度依赖性,行政机关有权依据强制披露的证据链,向整个技术供应链提起违约索赔或产品责任诉讼。

在追偿程序的具体设定上,应当在政务采购合同中确立按份比例与连带担保相结合的违约归责框架。行政机关在对外承担公法赔偿后,如何科学量化复杂链条上各技术主体的过错比例,是设定内部追偿机制的关键。为此,应当在政务应用系统中强制推行算法决策日志留存与场景复现制度。当侵权争议发生时,通过调取决策瞬间的系统日志,审计机构可以对致害结果进行逆向技术溯源,进而帮助法院审查决策是否存在合法性缺陷。^①

具体而言,如果技术复现结果表明,模型的违规输出是由于最终应用平台在部署时的参数微调失当或接口调用错误所致,则由该平台运营方承担主要赔偿责任。如果溯源发现偏见根源在于中间蒸馏环节进行了过度粗糙的模型压缩与剪枝,导致代表特定群体权益的特征权重被忽略,则负责蒸馏的中间开发者应承担核心违约责任。如果最终判定偏见深植于原始教师模型的预训练语料中且蒸馏过程不可避免地继承了该固有缺陷,原始模型的提供方即便未直接参与政务部署,也应依据政府采购的多方协议或技术担保条款承担相应的追偿份额。这种技术审计与合同法比例责任相嵌套的违约追偿机制,不仅打破了多轮蒸馏带来的责任隐匿技术屏障,避免了国家财政资金的流失,更能通过合同责任威慑,促使人工智能产业链上下游的开发者在模型设计的前端阶段主动融入合规审查程序,从而提升数字政务底层支持系统的公平性与安全性。

(三) 厘清政务模型应用中平台与开发者的协同合规义务

在数字政府的实际运行中,政务大模型的部署往往依托于第三方云平台或开放接口。此时,平台运营方与模型开发者的权责边界极易混淆。平台不仅提供算力支撑,还可能深度参与模型的微调与接口规划。为避免行政自动化决策违法,应在数字政务应用生态中厘清二者的多元主体合规义务。^②

首先,应将平台视为政务数据安全与模型合规的关键监督节点。在数字政务合作框架下,平台不仅是技术的承载者,而且是政务业务的延伸参与者。行政机关应通过政府采购协议与服务等级协议向承接政务模型的平台施加严格的审查与过滤义务。平台在将外部蒸馏模型接入政务专网前,必须核验其数据来源与算法备案证明。对于未能充分履行审查义务而导致不良模型引发政务风险的,行政机关可依据合同追究平台的违约责任,或将其从政务云供应商白名单中剔除。

其次,建立平台与开发者的风险隔离与信息传递机制。在处理涉及公共利益或个人隐私的政务数据时,平台应提供沙盒环境,确保模型开发者在进行线上微调时无法非法截留政务数据。同时,平台应当向行政机关承担算法透明度的辅助说明义务,当行政决策受到质疑时,平台需配合调取运行日志与调用记录,协助还原模型的决策轨迹。通过运用合同约束与政府供应商信用评价工具,将行政法的合法性要求内化为平台与开发者的商业合规准则,从而在不逾越行政权边界的前提下,实现对复杂技术生态的有效规制。

^① 参见王正鑫:《自动化行政裁量的司法审查》,载《比较法研究》2026年第1期,第211页。

^② 参见张凌寒:《〈个人信息保护法(草案)〉中的平台算法问责制及其完善》,载《经贸法律评论》2021年第1期,第45页。

五、结语

多轮知识蒸馏模型在数字政府建设中的广泛应用,在极大提升公共治理效能的同时,也深刻解构了传统行政法体系中建立在单一主体与线性因果之上的归责逻辑。面对由技术嵌套带来的责任隐匿与黑箱叠加问题,行政机关应立足于数字政务的公法属性,将监管重心向内部治理与采购合规转移。通过确立数字政务采购中的多轮蒸馏链条强制披露机制,并辅以建立在合同违约与信用监管之上的责任联动体系,行政机关方能在技术供应链中精准锚定过错节点。这不仅使行政主体在对外承担国家赔偿责任后能够实现有效的内部追偿,更从源头上防范了行政裁量过度依赖算法的深层风险。

人工智能治理已超越了私法范畴的个体纠纷解决,成为国家治理体系和治理能力现代化的重要命题,行政法理论在数字时代也面临如何进行自我革新的迫切问题。面对人机协同的复杂现实,对传统行政法中的行政主体、可归因性与行政公开等基础概念亟须进行再概念化诠释。行政机关必须实现从被动使用技术向主动规制技术的角色转型,通过提升跨学科的审查能力与技术穿透力,确保算法工具的运行轨迹始终被约束在法治政府的框架之内。

生成式政务大模型的行政法治理仍是一项长期且开放的系统性工程。随着智能化决策向行政裁决、行政执法等核心权力领域的深度渗透,如何重构数字时代的行政正当程序,特别是激活并本土化相对人的“自动化决策拒绝权”,将成为维护公民基本权利的关键防线。同时,针对算法歧视在公共服务中可能引发的隐蔽性不公,行政救济机制的适应性改造亦刻不容缓。在全球人工智能治理的竞合态势下,中国应积极汲取监管沙盒等前沿治理经验,将公平、透明、可问责等伦理准则转化为具有实质约束力的行政规范,努力在激发前沿技术创新活力与坚守法治核心价值之间探寻最优平衡,从而为数字中国建设贡献法治智慧。

The Administrative Law Liability Dilemma and Solutions in the Application of Multi-Round Distilled Large-Scale Models in Digital Government Construction

LI Xiansen

(School of Law, Beijing Institute of Technology, Beijing 100024, China)

Abstract: In the construction of digital government, multi-round knowledge distillation techniques have been widely applied in the private deployment of large-scale models. Compared with traditional algorithms, this technology's unique “teacher—intermediate student—final student” multi-level nested structure not only easily amplifies historical data biases but also causes the causal evidence chain to break during the intergenerational transmission of the model, leading to attribution dilemmas including difficulties in administrative entity confirmation, the failure to satisfy the legal burden of proof, and damage to administrative discretion rights. Facing complex chained models, the current usage-oriented and single-attribution regulatory logic is facing failure. To bridge the gap between technology and law, the regulatory path should not cross the bottom line of the rule of law to intervene in the front-end market, but should be based on the public law attribute of digital governance, and shift the regulatory focus to internal governance and procurement compliance. By establishing a mandatory disclosure mechanism for model chains in government procurement and complementing it with a contract-based liability linkage recourse mechanism, the fault nodes can be precisely anchored. This not only ensures that administrative entities can effectively seek recovery but also provides a legal and feasible institutional framework for artificial intelligence administrative governance.

Key words: multi-round knowledge distillation; application of large-scale models; attribution in administrative law; mandatory disclosure in government procurement; technical black box; linkage mechanism of liability

(责任编辑:沙晓岑)