算法黑箱的治理迷思与破解

袁 曾
（上海大学 法学院，上海 200444）

摘要：当前，有关人工智能的治理研究有脱离事实与逻辑的风险，其中，算法黑箱等研究及以此拓展的治理范式尤为突出。从技术层面分析，算法无非是数据驱动的复杂数学公式，脱离人工智能产品而仅就算法自身进行规制的立论并不客观。从治理层面分析，执着于打开算法黑箱只能造成计算系统概率性输出与法律因果关系间的责任链割裂，无法有效解决现实治理难题。从发展层面分析，算法作为知识密集型的智力劳动成果，应受知识产权法律的刚性保护，技术秘密强制公开将抑制创新活力。亟待破除以算法黑箱等为代表的人工智能治理迷思，从实际出发建立“整体可控制”的法治框架，实现数字法学研究从虚幻理论到有效可控的范式转变。

关键词：算法黑箱；算法解释；无人驾驶；新质生产力

中图分类号：D922.16 **文献标志码**：A **文章编号**：2096-028X（2025）04-0022-10

算法黑箱原指人工智能系统中输入与输出之间的隐层机制，难以被外界观察和理解，导致决策过程缺乏透明性，可能引发数据不可控、结果与事实相悖等问题，并形成信息泄露、劳动控制、就业歧视等风险。① 算法已经高度嵌入了人类生产与生活之中，特别是依赖其运行的大语言模型等生成式人工智能已大幅提升了机器的感知、理解与创造能力，使之无限逼近甚至超越人类的认知水平。基于此，有观点提出算法黑箱已对传统法律体系与制度构成巨大挑战，有必要在治理层面构建算法备案、解释、审查、测试、监管等复杂法律规制，通过明确责任划分的方式实现算法透明、公开，从而平衡数字技术发展与权益保护。② 客观来说，高速迭代的海量先进算法已融入社会运行的各项关键环节，算法黑箱是计算机系统运行中必然存在的事实。但从治理角度分析，在深入审视算法黑箱概念本身及其在制度构建中的核心问题后，基于算法黑箱形成的研究基点及以此衍生的“算法透明”“算法规制”“算法解释”等规制框架，无法解决当前技术与产业发展中的现实问题。人工智能治理的核心逻辑并不在于算法等技术层面的开箱检视，现行法律足以满足当前技术所处的生产力水平下法律关系的调节需求。为解构算法黑箱的研究迷思，需要破除执着于技术复杂性表象的治理逻辑，从有效可控的法治逻辑与以用为先的实践效能角度，准确界定算法乃至人工智能技术整体的规制内涵与框架，将数字法学等新兴学科的研究范式与促进算法、数据、算力等智能生产力要素的价值释放紧密结合。

一、算法黑箱及其衍生治理理论

算法黑箱是算法决策过程中技术性隐匿与法律认知结构性断裂的结果，被视为滋生算法偏见、歧视、操纵、共谋、垄断及信息滥用等社会问题的重要成因。③ 面对其已经或者可能引起的系统性风险，打开算法黑箱、寻求算法控制者等主体、提供算法解释或公开决策过程等观点，已成为当前法律规制与技术伦理研究的主流，并基于此形成了相应治理理论。④ 目前，“中国知网”中以法律规制为论述角度对算法黑箱进行研究的论文超过90篇，⑤主要分为以下三种流派。

基金项目：2024年度国家社科基金重大项目（24VHQ014）

作者简介：袁曾，男，法学博士，上海大学法学院副研究员，上海大学马克思主义研究院特聘研究员。

① 参见金雪涛：《算法治理：体系建构与措施进路》，载《人民论坛·学术前沿》2022年第10期，第49页。

② 参见班小辉：《论平台用工算法透明的制度实现》，载《清华法学》2025年第1期，第197-198页。

③ 参见张海柱：《行动者网络理论视域下的算法黑箱与风险治理》，载《科学学研究》2023年第9期，第1546页。

④ 参见解正山：《算法决策规制——以算法“解释权”为中心》，载《现代法学》2020年第1期，第184页。

⑤ 相关数据查询日期截至2025年8月7日。

一是“算法透明”论。由于算法的运行过程无法被观察且调用的数据、信息可能无法验证来源，因此有必要对算法的运算机理与数据调用等关键流程予以公开，以增强算法的合法运用与合规设计，尽可能地避免算法歧视、算法共谋。① “困在算法里的骑手”“热搜霸榜”等社会关注度极高的热点问题，均因算法运作的机械化与不透明化而导致劳动者或消费者被算法的控制者操控，因此有必要通过打开算法黑箱使对算法的治理成为可能。② 这类观点得到了部分回应，例如，《中央网络安全和信息化委员会办公室秘书局、工业和信息化部办公厅、公安部办公厅、国家市场监督管理总局办公厅关于开展“清朗·网络平台算法典型问题治理”专项行动的通知》针对侵害外卖骑手等新就业形态劳动者权益问题，提出“详细公示时间预估、费用计算、路线规划等算法规则”；针对互联网社交平台等存在的操纵“热搜”等重点问题，明确要求“全面公示热搜榜单算法原理，提升榜单透明度和可解释性”等。

二是“算法规制”论。人工智能的法律治理应尽可能深入认识和准确了解各式各样的人工智能系统，并使此种认识和理解尽可能得以动态、科学、量化的实现。③ 算法黑箱本质上是工程实践中的技术封存手段，其初衷是通过封装和抽象来管理日益复杂的算法系统，从而提升开发效率、保护核心逻辑、降低使用门槛。从技术层面分析，算法作为一种工具，理论上可以受到也应当受到监管。因此，有观点认为应着重针对算法加强监管，例如曾广泛被提及的“算法分类分级监管”“分层规制”“监管沙盒”等观点，甚至包括“算法审计”等创新论点。④ “算法规制”论的主要拥趸者提出，在当代法律体系“规范—行为”的框架之中，其逻辑起点在于以规范解释行为、以规范引导行为、以规范调整行为、以规范限定行为，进而将人的行为框定在规范体系内，因此，将基于合法性判断的规范规制转变为致力于算法可信的分层规制，将诸算法具有内在一致性的诸原则、诸方式、诸路径加以统一规制，以在治理层面实现对算法的有效控制。⑤

三是“算法解释”论。由于算法技术上的“不透明性”可以被“问题化”，有学者认为其阻碍了法律责任的清晰认定，即当算法决策导致损害时，难以追溯原因、划分责任主体，进而影响了实践中的救济与问责。特别是传统法学框架在涉及算法等新技术运用时，存在归责主体认定上的困境，现有知识体系难以解释由于技术构建产生的新问题，由此催生了“算法解释”论。⑥ 该流派要求控制者解释算法运行与输出的逻辑过程，进而明确算法责任主体与责任分配。⑦ 当算法控制者无法说明其算法运作的过程时，即应承担包括侵权责任在内的相应责任。从法律法规与政策实施层面分析，“算法解释”的立论构想亦得到了部分支持，例如，2021年《新一代人工智能伦理规范》特别强调在算法设计、实现、应用等环节提升透明性、可解释性。

有学者将算法黑箱分为三类：一是技术黑箱，即由于算法自主深度学习导致透明度缺失形成黑箱；二是组织黑箱，由于算法的使用者或应用平台不公开调用的数据等相关信息，使算法在应用层面缺失透明度；三是解释黑箱，由于算法具有突出的领域性与专业性，对于非专业群体而言算法本身就是难以理解的。⑧ 算法黑箱已成为算法透明度缺失及由此引起的各类风险问题的集成化与具象化体现，前述观点基于算法技术本身而提出了相应的解决办法与相对系统化的治理理论，从法学研究应为社会发展服务的角度分析应属无可非议，但仍具有无法回避的理论缺陷与实践桎梏。

二、算法黑箱治理理论的先天缺陷

（一）“算法透明”论与知识产权有效保护存在现实冲突

从技术的本质分析，算法黑箱仅是一种客观存在。当代知识产权法体系主要保护的法益是创新的智力

① 参见钟晓雯：《算法推荐网络服务提供者的权力异化及法律规制》，载《中国海商法研究》2022 年第 4 期，第 63-72 页；赵泽睿：《算法论证程序的意义——对法律规制算法的另一种思考》，载《中国政法大学学报》2023 年第 1 期，第 181-200 页。

② 参见吴勇：《超越劳动关系：平台用工算法管理规制的三重维度》，载《中国应用法学》2025 年第 2 期，第 51-66 页；周子羽：《算法行政“保险丝”机制：一个综合性算法规制分析框架》，载《探索与争鸣》2025 年第 2 期，第 180-181 页。

③ 参见苏宇：《从算法解释到系统测评——人工智能法治的信息工具变革》，载《探索与争鸣》2025 年第 3 期，第 116 页。

④ 参见蔡星月：《算法规制：从规范规制到分层规制》，载《现代法学》2024 年第 4 期，第 178-179 页。

⑤ 参见狄行思：《金融领域人工智能治理的公平进阶：算法可解释性与协同规制进路》，载《南京社会科学》2025 年第 7 期，第 88-90 页。

⑥ 参见陈京春：《算法的可解释性与刑法归责》，载《法律科学》2025 年第 4 期，第 12-14 页。

⑦ 参见费凯学：《算法辅助法律解释的规范性构建》，载《中国海商法研究》2024 年第 4 期，第 55-57 页；王仲羊：《刑事司法中算法解释的制度选择》，载《中国政法大学学报》2024 年第 6 期，第 258-261 页。

⑧ Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p.2-9.

成果，通过撰写公式、用系统的方法描述解决问题的策略形成的算法，因具有创新性和独特性，所以是各国知识产权法应重点保护的对象。《关于加强互联网信息服务算法综合治理的指导意见》明确指出，要坚持技术创新，大力推进算法创新研究工作，保护算法知识产权，提升算法的核心竞争力。有学者提出，算法符合商业秘密的法定构成要件，包括非公开性、商业价值性、保密措施。① 知识产权所保护的算法，是以计算机语言编译，以代码化指令序列为表现形式，由计算机运行并产生独立理性价值结果的程序算法，应突破传统范式并参考管制性排他权，构建针对算法核心特征的专门保护制度。② “算法透明”论提出了一种看似合理但实际上与实践矛盾的方案，在现行法律框架下，知识产权作为一种法定的排他性财产权，其保护具有优先性和刚性，对算法的透明度诉求与知识产权专有性、保密性存在必然的根本性冲突。这种冲突不是技术层面的展示困难，而是根植于法律权利设定本身的刚性壁垒。要求算法提供者进行可能使其丧失知识产权保护或承受巨大商业风险的深度公开，既缺乏充分的法律依据，也严重违背知识产权法激励创新的立法宗旨。

在尊重现有知识产权制度的前提下，打开算法黑箱的法律诉求注定无法取得较好的实践效果。算法领域的核心竞争力在于持续的创新迭代、优化的模型架构、独特的训练方法和精调的核心参数，这些既是知识产权保护的核心对象，也是激励创新的源泉。算法源代码的公开并不足以证明决策流程的公平性，反因算法的透明导致信息泄露，不仅使系统面临安全风险，还可能抑制算法创新活力。③ 例如，基于算法公开实现算法合谋，若平台经营者统一供应商的定价算法，随着算法监控市场与价格变动的能力不断提高，经营者可通过公开的算法预测其他经营者的价格，从而进行反向市场垄断。④ 围绕以风险治理为核心的新兴规制需求，欧盟对于算法透明度的要求主要适配于风险的整体治理。2023年4月18日，欧盟委员会正式运行欧洲算法透明度中心，汇集数据科学家、人工智能专家、社会科学家和法律专家，形成跨学科研究网络，以支持已出台的欧盟《数字服务法案》。2023年12月通过的欧盟《人工智能法案》被较多数字法学研究者视为较为成功的新兴科学技术立法范本，其第13条第1款规定，高风险人工智能系统的设计和开发应确保其运行具有足够的透明度，以便部署者能够解读系统的输出并恰当使用。尽管欧盟在立法技术上的先进性使其法律文本看似极为合理，但其周延的立论说明亦无法有效减缓欧洲自主算法技术的发展颓势，其中最为重要的原因在于强制性、深层次的算法透明度要求，必然与现行知识产权法律制度的根本原则和核心保护机制发生不可调和的冲突。算法早已是高度知识密集型的机制集成，无论是短视频推荐机制流程，抑或是资本市场走势预测，均需要开发者或控制者耗费大量的资金与技术投入，方能形成可供市场使用的智力成果。从经济学投资预期与回报产生的基本原理着手，相关利益主体天然期望获得知识产权的法律保护并取得相关的各类收益（即采取“闭源”策略）。强制公开虽然取得了透明的效果，但对于鼓励资本不断加大算法技术投入与取得正回馈并无益处，长期坚持“算法透明”论对于算法乃至人工智能整体的发展具有较大风险。

（二）“算法规制”论在规制实践中效能不足

就学术研究与实践发展而论，算法黑箱以及规制算法及其研发者、控制者等主体的观点已并非罕见。⑤ 其中，成体系性立法最为迅速的就是欧盟，囿于其数字技术实力的劣后地位与互联网经济的相对孱弱，欧盟期冀通过规则引领，重塑其在数字时代的领导地位。欧盟《人工智能法案》创造性地依据社会危害烈度、权利侵害不可逆性、系统性风险扩散力三项指标，将人工智能系统划分为低风险、高风险、不可接受风险、特定风险四类，以风险评级与控制作为对包括算法技术在内的人工智能规制内核，对人工智能实施分类分级的全面监管。该法案同时根据技术标准的成熟度，设定了清晰的分阶段实施路径。尽管该法案试图通过风险分层实现精准治理，但分类分级的规制方法始终存在无法回避的技术难题——即按照什么标准分类、按照什么类型分类、又按照何种层级与控制烈度进行分级管理？当前任何一家具备垄断实力的互联网企业均拥有海量的算法专利等先进技术且仍在不断快速升级，监管机构设定的分类分级能否真正追赶上技术迭代的速度，存在极大的不确定性。为了控制算法黑箱可能引起的法律问题而预留的风险敞口能否适时回应技术发展的

① 参见冯晓青、李可：《人工智能时代商业秘密保护规则重塑》，载《知识产权》2025年第5期，第27页。
② 参见靳雨露、肖尤丹：《算法的知识产权保护路径选择》，载《中国科学院院刊》2022年第10期，第1517页。
③ 参见[美]约叔华·A.克鲁尔、[美]乔安娜·休伊等：《可问责的算法》，沈伟伟、薛迪译，载《地方立法研究》2019年第4期，第113页。
④ 参见苏敏、夏杰长：《数字经济中竞争性垄断与算法合谋的治理困境》，载《财经问题研究》2021年第11期，第39页。
⑤ 参见郑琳：《反思与突破：自动驾驶地方先行立法的试验型进路》，载《中国海商法研究》2025年第1期，第91-102页。

现实需要,存在无法回避的结构性矛盾。同时,企业与研发者为了满足动态的分层分级监管将付出巨大的合规成本,直接干扰了先进技术发展所需的较优法治营商环境,企业不得不为应对不同风险等级而被迫建立复杂的对应流程。一方面,制定该类监管的标准极为复杂且进展缓慢,导致企业在合规建设过程中缺乏明确的标准指引,陷入发展的巨大不确定性中。另一方面,通过率先确立高风险系统禁用清单和生成内容标识规则,该法案迫使跨国企业被动适配其标准,使企业为满足监管需要付出的合规成本显著增加。该法案根据违法行为的严重程度和企业规模大小,设计了阶梯状的处罚条款。违法企业面临最高占其全球营业额 7%的罚款成本,但最关键的是企业可能在采取技术创新时并不知道其违反了何种法律与义务。①

“算法规制”论的核心治理目标是有效管控算法系统引发的整体性、系统性风险。② 但算法整体风险并非单一源于黑箱内部,而是多重因素在复杂系统中动态耦合、演化的结果,其风险贯穿于数据采集、模型构建与输出应用的全链条和全周期,呈现出显著的整体性、弥散性与不可分割性。在输出的结果方面,算法所依赖的数据存在不可知的输入污染源,训练数据本身即是社会偏见、历史歧视、采集偏差的复杂混合物,数据间的隐性关联远超预设范畴,难以通过监管算法内部逻辑来追溯这些深埋于海量、异构数据中的偏见源流。③ 算法运算后生成的虚构或错误内容将进一步加剧数据污染的负面影响,对算法本身进行规制的方法论也无法预知或防止特定数据污染在特定场景下触发的有害输出。④ 加之算法所运用的各类模型工具已高度复杂化,造成对监管手段与能力的要求极高。同时,算法通过强化学习,持续自我调整优化,其决策逻辑在部署后已发生动态演变,初始设计文档或静态代码审查,均无法反映其当前状态。⑤ 关键风险往往源于系统层面的“涌现”行为,即个体组件简单规则相互作用下产生的、无法从单个组件推导出的复杂整体行为,监管算法的局部逻辑或单个决策,根本无法捕捉或预测这类全局性、涌现性的风险模式。⑥ 算法所蕴含的认知鸿沟与专业解释壁垒难以逾越,其内部决策逻辑是随着对训练数据的学习而改变的,处理大量特别是具有异构属性的数据会增加代码的复杂性,同时也需要使用内置于代码中的技术和装置来管理它,这加剧了对计算过程的认识模糊。⑦ 通过深入剖析算法风险的本质与培育创新的规制机制,对算法“分类分级”的监管诉求非但难以有效管控整体风险,反而可能显著抑制算法等技术的创新活力,并潜在地加剧系统脆弱性,陷入“欲控风险、反增风险”的规制悖论。

(三)“算法解释”论无法解决责任承担的关键问题

司法裁判在应对新兴技术带来的挑战时的经典策略是类比推理和渐进调适,力图将新兴事物纳入既有理论框架与认知范畴中调整。⑧ 面对算法黑箱引发的责任认定问题,最自然的理论预期是将算法决策过程类比为某种已知的法律关系或行为模式,并通过解释算法的内部运行逻辑以理解此类新事物,进而厘清责任,确定义务承担主体。这正是许多“算法解释”论者呼吁“解释权”的潜在逻辑基础——为类比提供认知素材。⑨ 法律的真谛在于经验而非逻辑,某种法学理论提出后能否被有效地应用于指导司法实践,应是判断该理论是否具备生命力的最大前提。但分析自 2017 年算法突破技术原点以来的裁判,众多“算法解释”论的观点与论证逻辑并未在司法实践中予以引证。究其原因,一是“算法解释”或“解释权”缺乏坚实的法律请求权基础与可操作性规则。欧盟《一般数据保护条例》中的“解释权”条款范围模糊、效力不明,且在实践中常被降格为“结果解释”或“系统功能描述”,并非真正意义上的算法内部逻辑揭示。⑩ 二是传统侵权法的证明

---

① 欧盟《人工智能法案》强制要求高风险系统进行本土化合规认证,并设立欧盟人工智能办公室作为中央监管机构,强化对核心基础设施的控制权。但由于极其严苛的监管机制与方法,较大可能形成“寒蝉效应”,阻碍算法等人工智能技术的长期创新与发展。参见廖秀健、史丽冬:《外部性视角下欧盟〈人工智能法案〉的国家安全风险分析》,载《科技管理研究》2025 年第 9 期,第 241 页。

② 参见魏娟玲:《统筹发展和安全的算法治理逻辑及机制创新》,载《山西大学学报(哲学社会科学版)》2025 年第 3 期,第 122 页。

③ 参见陈俊秀、徐玉琴:《人工智能大语言模型引发的数据污染风险及其规制路径》,载《大连理工大学学报(社会科学版)》2025 年第 4 期,第 58 页。

④ 参见和军、李江涛:《人工智能数据风险及其治理》,载《中国特色社会主义研究》2024 年第 6 期,第 46 页。

⑤ 参见郭全中:《技术演化与涌现风险:生成式人工智能的协同式敏捷治理体系研究》,载《编辑之友》2025 年第 4 期,第 50 页。

⑥ 参见崔铁军、李莎莎:《基于大模型与因素的人工智能能力涌现》,载《重庆理工大学学报(自然科学)》2025 年第 2 期,第 84 页。

⑦ 参见董春雨:《从机器认识的不透明性看人工智能的本质及其限度》,载《中国社会科学》2023 年第 5 期,第 158 页。

⑧ 参见孙跃:《数字经济时代算法司法治理的挑战及其应对——基于算法正义的裁判方法与多元共治》,载广州市法学会编:《法治论坛》2023 年第 2 辑,法律出版社 2023 年版,第 51 页。

⑨ 参见桑先军:《司法算法解释机制的系统构建与运行》,载《法治现代化研究》2025 年第 3 期,第 160 页。

⑩ 参见陈浩林:《民法视角下“算法解释权”的质疑与重构》,载《西部法学评论》2023 年第 3 期,第 121 页。

责任规则并未因算法决策而颠覆。被侵权者通常只需证明损害、损害与算法输出决策的因果关系及被告存在过错即可，而非要求算法控制者通过解释算法的内部机制以“自证清白”。理解黑箱内部的复杂公式转换过程，对于解决判断输入数据的合法性、输出结果的危害性、决策主体的注意意义与可预见性等法律问题既非充分也非必要。三是即使司法机关依职权要求算法研发者或者控制者等责任主体出示对于算法的解释，“由谁来解释”“解释到什么程度才算充分”“如何验证解释的真实性与准确性”等关键问题均无成熟的标准或技术共识，导致解释要求在实践中易沦为形式，且更易由此产生无休止的争议。对此，有学者提出，应采取更为简便的方式处理算法歧视等现实问题，例如，在算法自动化决策的场景下，主张产品责任的治理逻辑并适用无过错责任原则，通过法律的倾斜性保护以弥合技术与发展的隔阂。① 为有效解决算法黑箱引发的法律问题，治理核心不应放在透视黑箱内部或要求责任主体进行解释，而在于把握现实问题的核心争议点与构建外部问责机制，“算法解释”论在司法实践中，既非法律要求，亦非有效工具，更非必需路径。

法律责任认定的核心基石在于确立清晰、确定的因果关系链，即需要在事实因果关系成立的基础上，判断行为事件与损害之间的联系是否足够紧密、直接、可预见，使其在法律上具有可归责性。② “算法解释”论所期待的效果，是在给定算法情景下，通过控制已知的输入端去判断相对未知的输出端，让生成的结果符合人类的认知。③ 基于深刻的实际与法理分析，解决法律责任认定的核心诉求，无法且不必通过理解算法内部封装的黑箱逻辑来实现，追求“解释”本身即是注定无解的理论路径。从客观实际分析，算法作为公式序列的本质，已经消解了解释的必要性。要求解释一个复杂的数学函数为何输出特定值，其答案只能是另一组更抽象的数学符号或概率分布，无法转化为法律可接纳的归责理由。对于算法治理，需要的是可以确定责任主体进行追责的依据，而非纯粹的算法技术还原。虽然相关性在一定程度上可以强化因果关系，由此算法解释具有现实合理性，④但实际上算法作为数学公式与数据的被动产物，其“表达”并非自主意志的产物，而是设计者意图、训练数据选择、参数设定、部署环境等综合作用的结果。其构建者在算法生命全周期中将不由自主地带入基于自身利益追求的观念倾向、数据收集和选择的非客观性、数据本身的质量问题等，这都会造成大数据分析结果出现偏差。⑤ 概率值的高低是相对于训练数据和模型本身而言的，具有内在的情境依赖性和不确定性。同一输入在不同模型或不同数据背景下可能得出不同概率。这种差异导致算法运行逻辑无法直接转化为法律论证所需的正当理由。算法输出结果源于输入变量间复杂、非线性、且通常不可完全追溯的相互作用，很难精确分离出单一或少数几个因素对特定输出结果的决定性贡献，数据收集、数据预处理、建模等任何环节上的处理不当，均可能埋下隐患。⑥ 从算法推理过程的概率性分析，算法是建立在概率统计的梳理基础上，通过大数据分析输出特定场景下的行为反馈，并据此对未来事件或未知状态进行概率性预测。⑦算法的决策过程并非基于绝对真理或逻辑必然，而是基于对海量数据的统计学习，归纳出概率性关联模式。算法解释本质是类似“$y=f(x)$”的函数映射，是统计结果的相关性描述，事实上的因果关系并不适用于法律上的因果关系的认定。⑧ 基于数学公式本身的客观性与确定性，经过算法处理后输出的内容从概念到逻辑、从语义到语法都进行了彻底的转换，导致输入与输出之间存在巨大的“语义鸿沟”，如果强行溯源解释必然导致关键信息的系统性失真或彻底丢失。算法运行的效果难以预测，极小误差都可能产生巨大的负面影响，对算法的极强合理性的期待不具备现实可能性。⑨ 寄希望于“算法解释”化解算法黑箱的可能风险，在现行法律框架内与人工智能有效治理的核心目标背道而驰，甚至可能引发更深层次的系统性风险。因此，应当在彻底解析算法黑箱作为治理核心命题的基础上，重构替代性规制路径与方法。

---

① 参见杨帆：《信用评分算法治理：算法规制与产品责任的融通》，载《电子政务》2022年第11期，第33页。
② 参见李姗姗：《自动驾驶汽车侵权责任研究》，载《合作经济与科技》2022年第17期，第18页。
③ 参见方凌智、程坦、顾倩妮：《剖析人工智能：算法逻辑、应用边界和管理影响研究》，载《研究与发展管理》2025年第3期，第3页。
④ 参见苏宇：《算法解释制度的体系化构建》，载《东方法学》2024年第1期，第84页。
⑤ 参见黎四奇：《大数据相关性对法律因果关系的挑战》，载《法律科学》2025年第4期，第114页。
⑥ 参见刘东亮、闫玥蓉：《大数据分析中的相关性和因果关系》，载《国家检察官学院学报》2023年第2期，第36页。
⑦ 参见郑智航：《人工智能算法的伦理危机与法律规制》，载《法律科学》2021年第1期，第16页。
⑧ 参见季卫东：《法律与概率——不确定的世界与决策风险》，载《地方立法研究》2021年第1期，第16页。
⑨ 参见杜严勇：《智能社会建构中的算法文化：本质特征、伦理风险及其规避路径》，载《同济大学学报（社会科学版）》2023年第1期，第80-81页。

三、域外技术治理经验的批判借鉴

考虑算法黑箱及其可能引发的治理风险，应超越技术客观事实本身，更加注重算法黑箱争议背后所蕴含的技术霸权竞逐。当前世界各主要经济体均在加码包括算法在内的人工智能投入与布局，新科技时代法律的制定与修改应充分服务于技术发展与社会实践。如前所述，当前国际治理格局呈现显著分化，欧盟期冀以其《人工智能法案》《一般数据保护条例》等规则引领技术治理，构建了严苛的人工智能监管体系，追求人工智能风险的同步监管治理，通过制定风险分层、事前合规、高额处罚等周密规则体系预设技术发展轨道以实现技术驯服，虽在风险分类分层等具体规制模式上彰显了保护基本权利的决心，却饱含规制僵化与创新抑制之风险。反观美国，其秉持技术实用主义与市场优先理念，为保持优势地位，其联邦层面明确拒绝严苛的统一监管，倚重行业自律、标准制定与事后追责，其优势在于为创新留足空间，但碎片化、执行力弱的弊端亦可能纵容系统性风险积聚。美国政府将促进人工智能的创新和发展作为高度优先事项，极力减少人工智能开发和部署的不必要障碍，重点保护其技术、经济与核心价值观。

相较于欧盟《人工智能法案》的全面强制性立法，美国更强调创新驱动和企业发展，以防止各州的分散立法与差异监管损害人工智能创新的整体格局。2025 年 5 月，美国众议院通过《HR1 法案》，该法案第 43201(c)节规定："任何州或其政治分区不得在本法案通过后 10 年期间内，执行任何监管人工智能模型、人工智能系统或自动决策系统的法律或法规。"这一禁令覆盖从算法设计到系统部署的全链条，致使其各州现行人工智能相关监管措施暂时失效，进一步放松了对人工智能的创新监管。该法案明确设定为期十年的监管暂停期，消除了各州法规拼凑的复杂性，为算法研发、产品商业化提供稳定环境，其核心逻辑在于深刻认识到在技术爆炸性发展的初期，苛求算法在运行过程全程透明可能严重阻碍技术创新。这一制度设计本质上是对"算法黑箱必然导致不可控风险"预设的否定，监管的智慧在于为创新预留必要空间，相信市场与技术自身的迭代能力能够逐步解决透明性问题，而非在技术尚不成熟的阶段以黑箱为名进行过度干预。美国对包括算法在内的人工智能治理从未脱离其国家竞争战略。

美国政府通过简化规则、扩大技术输出，重新激活本土创新链条，同时以更强势的姿态主导全球 AI 标准与市场秩序，试图以技术优势巩固美国在全球 AI 领域的主导地位。对内，美国政府通过巨额投资引导、税收优惠、公私合作等强力措施，倾举国之力扶持本土企业研发尖端、自主可控的 AI 系统。诸多举措的核心关切在于技术的先进性、可控性与战略优势的维系，而非纠缠于算法是否完全透明或可被解释。只要最终的技术成果强大、可靠且服务于国家利益，内部机制的精妙复杂反而可能构成一种竞争优势而非缺陷。① 对外，美国政府抨击欧盟对美国科技公司的监管是"变相征税"，同时通过出口管制等手段系统性遏制中国在先进人工智能领域的发展。美国《爱国者法案》《云法案》为滥用长臂管辖、进行域外制裁提供了依据，全方位限制了中国在人工智能和高性能计算领域的发展。在国家核心竞争力的博弈场上，对算法黑箱的哲学式忧虑，远不及对现实技术优势的巩固与对潜在战略竞争对手的压制显得紧迫与实际。观察美国技术治理逻辑的本质，算法等议题在国家战略面前具有高度的可协商性与工具性，其重要性常让位于更为根本的国家利益与技术霸权考量。当然，这种治理范式并非不存在缺陷，美国倾向于技术优先的行业自律路径，其"轻事前监管、重事后追责"的模式导致各州已出台的涉及深度伪造、自动化招聘歧视等问题的多项法律被废除，亦有可能导致基本权利的保护失去有效屏障，因此需要对其予以批判吸收。

四、对算法治理理论的重构

算法黑箱作为技术恐惧的具象化身，已被视为横亘于技术与法治之间的认知鸿沟，近年来在数字法学等领域被反复论述，似已成为不证自明的治理前提。但算法黑箱本身所蕴含的认知迷思并无法有效解决现实问题，"算法透明""算法规制""算法解释"等治理理论过度聚焦于算法内部决策过程的不可透视性，却忽视了规制人工智能的法律整体效果。需要将治理范式从算法黑箱的理论研究转为支持、鼓励算法等先进技术

① 参见周冉：《美国科技公司崛起下的数字地缘变局：一种权力嵌合的分析》，载《国际安全研究》2025 年第 4 期，第 154 页。

发展的应用研究，使得数字法学研究真正成为更具法理正当性与实践效能的研究领域与研究方法。

（一）体系性治理算法风险

当前中国以算法规制作为人工智能治理的框架核心，以《算法推荐管理规定》《生成式人工智能服务管理办法》《关于构建数据基础制度更好发挥数据要素作用的意见》等为代表的治理实践，敏锐地抓住了算法与数据两大关键要素，但在具体规制方法上深陷要素拆分、局部规控的窠臼，未能跃升至系统整合、动态适应的发展型治理范式。生成式人工智能已经逐渐突破大语言模型的单一文本限制，朝着多模态模型发展，通过跨模态对齐技术实现文本、图像、视频的联合编码，逼近人类综合感知能力，各类生成模型越来越多地与商业、教育、媒体等日常生活行为融合，因此，引导构建可预期的清晰法律框架极有必要。基于前文对算法黑箱命题的剖析及对欧盟、美国立法经验的扬弃，中国人工智能治理的转向在于超越对单一技术要素的透明性苛求与碎片化约束，通过构建“整体可控制”的人工智能治理路径及以敏捷治理、风险适配、创新包容为原则的整体性规则体系，兼顾发展与风险之间的平衡，选择具有中国特色的立法技术路径，最终实现技术创新活力与安全保障效能的高阶平衡。

一是正视技术迭代现实与立法体系保持总体稳定的要求，明确技术发展应更好地回应公共利益诉求。[①] 需要摒弃“算法透明”“算法规制”等“一刀切”式或放任自流的规制方式，在原则和实用主义之间取得适当的平衡。在新科技立法上采取审慎推进的态度，若现有法律法规可以通过解释、修改等完成治理要求，则以不出台专门法律为主，最小化减少企业等主体的合规损耗。二是平衡风险不可预见性与权益保障明确性，建立预防性责任框架，以落实责任分配。通过投保新技术强制责任保险、建立赔偿基金等手段应对未知风险。以欧盟的合规治理为例，大型企业一旦违规将面临巨额罚款，而中小企业缺乏法律和技术资源以满足合规要求，可能导致企业暂停开发或创新活力降低的后果，甚至产生企业试图搬迁至规则更灵活的司法管辖区的现实后果。对于算法的治理，应避免欧盟高罚单、低清晰度的弊端，建立正向激励的精准治理机制，以平衡经济收益与企业责任风险。[②] 三是从总体国家安全观的高度，明确技术自主与国家安全边界，建构安全利用、有效利用的风险治理范式，强化底线思维、极限思维，有效防范、化解人工智能可能产生的系统性风险，而非算法所引起的单一风险。以加快发展新质生产力为主线，统筹新质基础设施建设，在算力、算法、数据等智能核心要素层面构建“以我为主、安全可控”的治理基座。[③] 以更专业的治理理论、方法与实践体系，建设优质、可信、负责任的人工智能应用环境，通过专业化资源建设导向促进高质量数据资源的建设。[④] 按照客观规律，对算力布局规划、数据转化利用等智能底座的可控发展，采取规章先行的模式予以一体化规制，将成熟经验逐步转化为法律。四是将技术创新的标准化规制转化为推动产业发展的实际动力。人工智能的治理范式受到各国的历史脉络、文化土壤、传统习俗与价值体系的影响，全球范围内呈现出“多中心、低协同”的特征，规则协调难度大。当前对包括算法在内的人工智能治理，国际上尚未建立起明确的历史路径依赖。人工智能本身具有技术不确定、概念模糊及创新性强的特点，这使得它成为通过治理叙事塑造政策框架和提升社会接受度的理想领域。[⑤] 技术标准本质是“法律规范的技术化延伸”，通过将伦理要求转化为可验证的技术参数，使软法规范逐步获得实体法的实际执行力。中国在技术方面已具备较强的综合实力与发展潜力，利用自身硬实力优势，通过产业链赋能将有效提升国内标准的国际影响力，而标准特别是简洁高效规制范式的泛在适用又将有力促进人工智能产业的整体发展。Deepseek、文心一言等国产人工智能产品的持续走强与均等化服务，为科技实力的战略投送与留白空间的标准权制定提供宝贵的市场机遇。为有效发挥非对称性优势，应摒弃对算法黑箱治理理论等碎片化的应对策略，立体构建具有域外效力的技术标准法律治理体系，以可控的产业优势拓展数字治理疆域。

（二）以结果导向修正算法黑箱理论

算法黑箱及其衍生的治理理论虽具备天然的弊病，但此类理论为了解决社会问题的初心无可非议，关键

① 参见高秦伟：《人工智能标准规制的监督机制》，载《郑州大学学报（哲学社会科学版）》2025年第3期，第49页。

② 参见袁曾：《生成式人工智能责任规制的法律问题研究》，载《法学杂志》2023年第4期，第126页。

③ 参见李海舰：《加快发展新质生产力：理论内涵与实践逻辑》，载《经济与管理》2025年第5期，第9页。

④ 参见周文泓、熊小芳、叶雅寒：《面向人工智能的数据治理框架研究——基于政策文本的构建与展望》，载《信息资源管理学报》2025年第4期，第13页。

⑤ 参见翟仲：《论军用无人潜航器的豁免权及其规制》，载《中国海商法研究》2023年第2期，第94-103页。

问题在于如何有效解决算法黑箱化运行引发的现实社会问题，其中，核心是如何在尽可能减少法律调整摩擦成本的预期下解决责任承担问题。为有效平衡风险与发展，现阶段最有效的规制范式是通过结果控制行为的可能导向，而非将所有算法纳入监管或解释范畴。例如，在关注度较高的劳动者权益保护领域，算法黑箱的监管可以被转化为互联网平台管理义务的履行判断，平台是否公开路径规划的算法代码与骑手劳动强度过高的争议并无实际关联，有关联的应是互联网平台是否有效恰当地履行了《中华人民共和国劳动法》（简称《劳动法》）下的劳动者保护义务。在司法实践中，面对外卖骑手因配送算法压缩时间引发的交通事故索赔纠纷，裁判机关也并未要求平台公开路径优化算法的代码逻辑，而是聚焦于“平台是否对高风险配送时限设置具备预见可能性”及“是否采取必要安全防护措施”等管理维度。① 将算法黑箱问题转化为组织义务的履行评价，重点考察算法歧视行为是否给消费者带来实际损害，可以印证司法实践对算法透明度审查的实践态度。② 在监管最为严苛的欧洲，外卖员、网约车司机等零工从业者针对平台算法的诉讼已陆续出现。在“Deliveroo 骑手案”中，意大利博洛尼亚法院面对平台通过算法对骑手进行信誉排名的争议，也并未要求平台公开算法源代码，而是提出算法会处罚所有缺勤骑手，且未区分无故缺勤和由于生病、罢工等无法出勤的原因。③ 法院基于欧盟《一般数据保护条例》第 22 条认定平台未建立人工干预机制的行为违反了自动化决策限制规定，构成对骑手的歧视。此判决的精妙之处在于将争议焦点从算法内部逻辑的不可知性，转向算法运行结果的管理正当性，将算法黑箱的问题回归于对实体结果的判断。2023 年 1 月 26 日，美国国家标准与技术研究院公布《人工智能风险管理框架》，将治理视野进一步扩展至算法全生命周期，以系统性风险管控替代对透明化的单一追求。通过指导组织机构在开发和部署人工智能系统时降低安全风险，避免产生偏见和其他负面后果，以提高人工智能可信度，实现保护公民权利的目的。

司法系统的运行逻辑与资源禀赋决定了其对治理工具的选择必然以可行性、效率性与法理正当性为基础，以结果回溯责任主体的确定具备实践效能。若对价格歧视、金融风控、平台责任、无人驾驶等凡是涉及算法的案件均要求详尽的技术解释与验证，从司法实践的角度无疑将导致案件审理周期呈指数级延长，消耗本已稀缺的司法资源。加之复杂算法解释往往涉及冗长的听证、反复的技术质证与专家对抗，显著抬高诉讼成本，使当事人陷入“技术维权马拉松”，实质阻碍司法救济的可及性。④ 特别是在解释的过程中，当法官需要依赖算法研发者、控制者所提供的技术解释并寻求专家证人确定责任归属时，实则将关键的事实认定权与证据解释权让渡给了其他技术主体。在法官无法理解复杂技术原理与逻辑的背景下，这种依赖关系很有可能干扰司法的中立性与独立性。⑤ 司法审查的重心应置于算法决策的结果合法性、系统运行的可靠性及人类主体的行为合规性，而非工具自身的特殊性。⑥

以实际损害结果为导向，回归侵权法、合同法等传统部门法的成熟归责逻辑，是较为现实的法治效率选择。⑦ 法律评价的重心应是算法应用所引发的具体侵害事实，无论算法决策过程如何难以预测或观察，损害结果的发生及其与行为人之间的因果关系，始终是责任判定的基石。当无人驾驶汽车因系统误判引发事故，法律的核心追问不应是算法当时如何思考，而是开发运营者是否尽到合理注意义务、系统设计是否存在可预见的缺陷、是否符合法定的安全标准。监管者足以根据生成式人工智能的生命周期，分析在模型训练阶段、设计阶段、输出阶段可能存在的“不合理危险”，即产品缺陷，进而确定发展风险抗辩适用的条件与标准。⑧ 以结果而非过程作为责任锚点，能够避免陷入对算法透明、规制或可解释等技术目标的无限追逐，反而消解了法律应当具备的评价与归责功能。深入观察涉及算法侵权的司法实践，现有案件的裁判者均未执着于要求打开黑箱或要求解释算法的内部逻辑。以广受关注的“TikTok 程序算法推荐案”为例，上诉法院认定，当

① 参见罗智敏：《算法歧视的司法审查——意大利户户送有限责任公司算法歧视案评析》，载《交大法学》2021 年第 2 期，第 187 页。

② 参见李丹：《论算法歧视消费者的侵权责任认定——基于司法裁判的实证考察》，载《当代法学》2023 年第 6 期，第 82 页。

③ Ilaria Purificato, *Behind the Scenes of Deliveroo' s Algorithm: The Discriminatory Effect of Frank' s Blindness*, Italian Labour Law E-Journal, Vol. 14:169, p.169-194(2021).

④ 参见班小辉：《论平台用工算法透明的制度实现》，载《清华法学》2025 年第 1 期，第 198 页。

⑤ 参见袁曾：《生成式人工智能治理的法律回应》，载《上海大学学报（社会科学版）》2024 年第 1 期，第 30 页。

⑥ 参见曾迪：《算法价格歧视违法性认定的挑战与应对》，载《中国流通经济》2025 年第 2 期，第 112 页。

⑦ 参见周江伟、赵瑜：《人工智能治理原则的实践导向：可靠性、问责制与社会协同》，载《治理研究》2023 年第 5 期，第 119 页。

⑧ 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 80 页。

算法推荐构成平台自主表达时，不能再援引技术中立原则免责。[①] 此类域外判决跳出了“须理解‘黑箱’内部构造才能进行法律评价”的理论桎梏，将法律对技术评价的锚点定位于算法运行的客观输出与可观测的社会效果之中，有效规避了算法黑箱带来的治理僵局，有力印证了过程黑箱并不等同于结果失控或责任真空。

（三）以“技术可控制”取代“算法可解释”

“算法解释”论等作为技术工具治理理论，只能在相关性层面提供信息，无法在必然性层面建立起法律所需的、稳固的、个体化的因果关系认定。执着于通过算法解释来理解内部机制以认定责任，是人工智能治理方法论上的严重错位。当前学者所提出的算法解释路径，无论是“全局可解释”还是“局部可解释”，本质都是对复杂模型行为的近似模拟或归因分配，其本身也是一种概率性的、可能不稳定的解释模型，无法揭示算法决策的真正内涵。[②] 算法的公式本质决定了解释只能是事后合理化。[③] 算法不能成为责任黑洞，将主要责任都归因于算法，不但在实践中无法真正地回应和解决算法歧视问题，还易使得歧视行为的实施者隐藏在技术屏障后并逃脱制裁。[④] 将打开算法黑箱作为解决算法争议的核心路径，不仅因技术复杂性而难以实现，更可能扭曲现行包括产品责任在内的责任认定逻辑，使得算法的研发者、控制者甚至使用者对于自身在何种情况下可能承担责任感到无所适从，最终背离算法治理的初衷。当技术复杂性超越人类认知边界时，算法等深度学习模型的非线性决策本质决定其无法提供符合法律期待的可理解因果链，所谓解释往往沦为事后的合理性重构，既无法验证真实性，亦难保障决策公正。应将“算法解释”论的立论基础，逐步修正为算法技术等可控制的责任调整路径，确保人工智能发展的“整体可控制”。[⑤]

在规制内容上，应当以可视化风险作为规制基础。算法风险高度依赖具体部署场景、用户交互、环境变化，黑箱被打开后进行解释的内部逻辑，通常是静态的、脱离具体语境的。对算法内部逻辑的直接体现或解释，无法满足外部的、情境化的价值冲突，输出结果的“可接受性”远超代码的可解释范畴。[⑥] 过度依赖和满足于算法可以被解释的假象，可能导致监管者和开发者忽视真正关键但难以解释的整体性风险，资源被错误地导向审查微观逻辑的透明程度上，而非投放于压力测试、影响评估、实时监控、故障应急预案等能真正提升系统韧性的措施。向公众提供难以理解甚至被操纵的解释，可能制造一种虚假的“可控感”和“透明度信任”，一旦发生由系统性风险或不可解释因素导致的重大事故，这种信任将瞬间崩塌，从而引发更严重的信任危机和社会反弹。应超越对算法内部机理的原理解释，转而以总体国家安全观为统领，以外部审计替代内部解释，对人工智能应用引发的系统性风险进行精准识别与统筹治理。[⑦] 算法造成潜在威胁的本质不在于其决策逻辑是否可解，而在于其部署与应用是否可能危害国家安全、社会秩序、公民权利等重大法益。应将人工智能治理与实务相结合，依据《中华人民共和国个人信息保护法》《中华人民共和国反垄断法》等法律保护法益，对算法可能引发的歧视性风险、安全性风险及剥削性风险进行动态监测。[⑧] 以人脸识别技术为例，其核心风险点并非算法本身的黑箱特性，而在于大规模生物信息采集带来的隐私泄露风险、歧视性识别结果造成的社会不公、监控能力滥用对公民自由的潜在压制，因此，可基于风险控制构建个人信息保护路径。[⑨]《生成式人工智能服务管理暂行办法》即体现了此种风险导向思维，其核心关切在于服务提供者的主体责任、数据安全、内容合规及用户权益保障等宏观风险维度，而非强制穿透具体算法逻辑。

在归责方法上，需要穿透算法背后的控制主体。从更深层次分析，算法运行的结果绝非无源之水或无本之木，其本质是背后操控者即平台或企业的意志与价值取向的延伸与具象化。算法决策的偏好设定、目标函数的选择、数据源的筛选与标注，无不深刻烙刻着设计者和应用主体的意图与选择。以外卖平台配送算法为例，其催生困在系统里的骑手与频频发生的交通风险，算法不过是企业意志的执行工具，真正的黑箱是隐藏

① *Anderson v. TikTok Inc*, No.22-3061(3d Cir. 2024).
② 参见杨志航：《算法透明实现的另一种可能：可解释人工智能》，载《行政法学研究》2024年第3期，第159页。
③ 参见万方：《算法治理应聚焦解决的关键性问题》，载《理论探索》2022年第2期，第124页。
④ 参见王淑瑶、张钦昱：《算法歧视的理论反思与算法决策的规范重构》，载《电子政务》2024年第10期，第110页。
⑤ 参见袁曾：《算法应当被解释吗？——人工智能“可控制”的治理向度》，载《法学论坛》2025年第1期，第139页。
⑥ 参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，载《环球法律评论》2019年第6期，第29页。
⑦ 参见张永忠、张宝山：《算法规制的路径创新：论我国算法审计制度的构建》，载《电子政务》2022年第10期，第50页。
⑧ 参见敦帅、陈强、贾婷：《中国人工智能治理研究述评与展望》，载《中国科技论坛》2025年第4期，第41页。
⑨ 参见斜晓东：《风险与控制：论生成式人工智能应用的个人信息保护》，载《政法论丛》2023年第4期，第64页。

在代码与公式背后的商业伦理缺失。[①] 平台企业将极致效率与成本压缩奉为圭臬，通过算法规则的设计，如严苛的送达时限、不断优化的最短路径计算、高额超时惩罚机制予以刚性传导。其本质是利用算法工具实施劳动力压榨，此时，法律应当审查“30 分钟送达”等配送标准本身是否违反《劳动法》关于合理劳动强度的强制性规定。法律规制的矛头必须精准指向算法背后的责任主体，要求其承担与其技术能力、市场地位相匹配的更高注意义务与社会责任，矫正其可能失衡的价值追求，确保算法承载的利益分配符合法律设定的公平正义框架。

五、结语

人类对技术治理的认知始终随着生产力的进步动态演进，以“算法透明”“算法规制”“算法解释”为核心的传统规制范式，在生成式人工智能从大语言模型向多模态跃升的过程中已经陷入规制被动。有关算法黑箱治理的讨论本质上是技术封存体法律化表述的命题，其既受限于数学函数集的概率性输出与法定因果归责的错位，更面临知识产权保护刚性要求与司法实践理性回避的多重限制。算法的概率性、相对性、多因素耦合性本质，与法律对确定性、个体化、充分因果链的追求，存在根本性的、本体论层面的巨大鸿沟，导致算法黑箱在人工智能技术治理中面临深层适配困难，法律责任的认定需清晰连接行为主体、因果关系、损害结果，即使强行打开黑箱，得到的仍是复杂的相关性权重，而非法律可识别的、具有可归责性的行为及其与损害之间的规范性因果关系。当前人工智能迭代速度远超治理规则更新进程，面对“技术达尔文主义”的挑战，需要以稳定的规范与明确的责任预期促进技术的发展，在尽可能减少立法调整产生的摩擦成本基础上，通过法治手段破解风险与责任的矛盾，通过制度设计将技术治理转化为发展动能。亟需超越当前算法治理的理论路径依赖，实现人工智能治理从认知迷思向实质可控的现实跨越，重构多维可控的治理维度。从算力、算法、数据等核心要素统筹规划，考虑与投资激励、风险分配、责任承担比例适配的有效法治框架，通过建构以“可控制”为内核的体系化规制范式，创设包容技术创新的制度环境，以制度的韧性、包容性与前瞻性容纳技术的复杂性与不确定性，确保包括算法在内的人工智能技术高效服务于人类福祉与社会进步。

The Governance Paradox and Solution of the Algorithm Black Box

YUAN Zeng

(Law School, Shanghai University, Shanghai 200444, China)

Abstract: Current research on the governance of artificial intelligence runs the risk of detaching from facts and logic, with studies on the algorithm black box and the governance paradigms expanded by them being particularly prominent. From a technical perspective, algorithms are merely complex mathematical formulas driven by data. The argument that regulation should be imposed on algorithms alone, divorced from the context of artificial intelligence products, is not objective. From a governance perspective, insisting on opening the algorithm black box can only lead to a break in the chain of responsibility between the probabilistic output of computing systems and legal causality, and cannot effectively solve real governance problems. From a development perspective, algorithms, as knowledge-intensive intellectual labor achievements, should be protected by the rigidity of intellectual property laws. Compulsory disclosure of technical secrets will suppress innovation vitality. It is urgent to dispel the governance myths of artificial intelligence represented by the algorithm black box and to establish a holistically controllable legal framework based on reality, thereby achieving a paradigm shift in digital legal studies from illusory theories to effective and controllable ones.

Key words: algorithm black box; algorithm explanation; driverless vehicles; new quality productive forces

① 参见袁曾：《数字法学研究现状的再反思——法学理论向何处去?》，载《上海政法学院学报》2023 年第 3 期，第 122-124 页。