从算法“黑箱”走向算法透明：基于“硬法—软法”的二元法治理模式

钟晓雯

（华南农业大学 人文与法学学院，广东 广州 510642）

摘要：算法是“程序设计+数学规则”的集合，其“黑箱”特性虽然并非算法风险的唯一原因，但却是规制算法风险必须要解决的问题。强调开放与控制并重的“硬法—软法”范式能够为算法“黑箱”的治理提供一种新的立法框架。基于理论层面，硬法规范的初次外在规制决定了算法“黑箱”治理的裁量边界，软法规范的二次内在规制能够为政府的算法“黑箱”治理行为开放公众参与机制和提供内在参照标准。基于实践层面，欧盟与美国已形成的算法软硬法混合治理格局能够为中国建立算法“黑箱”的“硬法—软法”二元法治理模式提供借鉴。算法“黑箱”治理的软法规范应包括增强算法模型的可解释性和向社会公开算法源代码，同时，宜借鉴“遵守或解释”机制，要求算法开发者/运营者在不遵守软法规范时应作出合理解释说明。算法“黑箱”治理的硬法规范应包括算法权利（最为密切的是算法解释权、算法自动化决策拒绝权），以及借助算法备案、影响评估和合规审计启动算法问责。

关键词：算法“黑箱”；算法透明；算法可解释性；算法开源；算法权利；算法问责

中图分类号：D922.16　　**文献标志码**：A　　**文章编号**：2096-028X（2023）04-0053-10

From Algorithmic “Black Box” to Algorithmic Transparency: A Dualistic Law Governance Model Based on “Hard Law-Soft Law”

ZHONG Xiaowen

（School of Humanities and Law, South China Agricultural University, Guangzhou 510642, China）

Abstract: Algorithm is a collection of “programming+mathematical rules”. Although its “black-box” nature is not the only cause of algorithmic risk, it is an issue that must be addressed when regulating algorithmic risk. The “hard law-soft law” paradigm, which emphasizes both openness and control, can provide a new legislative framework for the governance of algorithmic “black box”. From the theoretical perspective, the primary external regulation of hard law norms determines the discretionary boundary of algorithmic “black box” governance, while the secondary internal regulation of soft law norms opens up the public participation mechanism and provides internal reference standards for the government’s algorithmic “black box” governance. From the practical perspective, the European Union and the United States have established a pattern of mixed hard law and soft law governance of algorithm, which can provide a reference for the establishment of “hard law-soft law” regulatory model of algorithmic “black box” in China. Soft law regulation of algorithmic “black box” should include enhancing the interpretability of algorithmic models and disclosing the source code of algorithm to the public, and at the same time, it is appropriate to draw on the “comply-or-explain” mechanism, which requires algorithmic developers/operators to provide reasonable explanations in the event of non-compliance with the soft law regulation. Hard law regulation of algorithmic “black box” should include algorithmic rights (the right to interpret algorithm and the right to reject automated algorithmic decision-making which are most closely related), as well as algorithmic accountability through algorithmic filings, impact assessments and compliance audits.

Key words: algorithmic “black box”; algorithmic transparency; algorithmic interpretability; disclosure of the algorithmic source code; algorithmic rights; algorithmic accountability

收稿日期：2023-09-21

基金项目：2023年度教育部规划基金项目“基于文本卷积神经网络的司法判决预测的实现路径研究”（23YJAZH004）

作者简介：钟晓雯，女，华南农业大学人文与法学学院讲师。

一、问题的提出

当前已进入以人工智能技术为核心驱动力的第四次科技革命时代,人类社会也渐趋从网络社会转向以算法为主导的"计算社会",算法的核心地位愈发凸显。从技术层面来看,算法是一种在数据采集与训练的基础上,依据特定的运算规则输出结果以完成目标任务的计算机程序,是"程序设计+数学规则"的集合。算法的运行包括"数据输入—模型运算—决策输出"三个步骤,但人们至多可以观察到算法的数据输入与决策输出两个步骤,无法完全掌握算法的内部运算或决策程序,于是便形成了算法"黑箱"。"黑箱"是关于"不透明"的一个隐喻,其原本是控制论的概念,指的是"只能得到它的输入值和输出值,而不知道其内部结构"①的系统。一般认为,算法"黑箱"的形成原因有三种:一是公司或国家保密,即公司或国家为保护算法而导致的不透明,如视算法为商业秘密;二是技术文盲(technical illiteracy),即机器学习算法与人类认知存在差异,人类难以理解;三是技术"黑箱",即机器学习的技术原理所固有的"黑箱"性质。②

之所以用"黑箱"形容算法不透明的特性,不仅是因为它符合控制论的"黑箱"特征,更在于它带给人类的潜在风险乃至社会恐惧。算法"黑箱"最直接的表现就是极易侵犯用户的知情权和个人隐私。算法"黑箱"使得诸多算法流程处于不透明状态,用户无法完全掌握算法的设计意图、内部运算以及决策程序等,这种信息不对称和不公开导致用户无法从外部直接观察和验证算法的数据收集、分析和挖掘行为,即便用户个人隐私在这些算法行为中受到侵害也难以察觉。除社会个体外,政府公共部门也难以回避算法"黑箱"带来的挑战。数字化时代的资源核心是信息技术,掌握了信息技术的主体在社会中更能拥有支配力,信息技术的归属有可能会打破现有的权力分配结构和社会秩序。③ 如此一来,掌握并操纵算法的私人技术公司实际上凭借这种技术资源在社会运行秩序中占据了优势地位,逐渐取得了对其他社会主体乃至政府公共部门的支配力。相反,政府公共部门逐渐失去了算法监督权和控制权,面临着被"算法社会"边缘化和去中心化的挑战。④ 总的来说,随着算法广泛、深度的融合应用,算法"黑箱"往往与算法歧视、算法垄断、算法合谋等负面评价绑定在一起,并逐渐形成一种隐性规训和统治空间:算法作为数字社会的主宰,借助"黑箱"效应伸出无形之手,逐渐架空人类在现实世界中的主体地位并控制人们的生活。因此,算法"黑箱"虽然并非算法风险的唯一原因,但却是规制算法风险必须要解决的问题。

乌尔里希·贝克(Ulric Beck)的社会风险理论指出,技术不仅是推动人类步入风险社会的动因,也是风险社会的主要表征。⑤ 现代对于诸多因技术发展引起的问题需要在风险社会的逻辑背景下探究其法律规制理念和路径。在"计算社会"下,因算法"黑箱"具有不确定性与不可预知性、损害后果具有累积性与扩散性等复杂情景,人类不得不穷尽可能手段对其进行治理。但一方面,算法"黑箱"衍生的社会风险是一种新型的社会风险,穷尽中国既有法律制度也难以充分治理;另一方面,传统政府治理遵循的是对抗、威慑和法律服从的制度逻辑,在治理算法"黑箱"时会陷入两难境地:或因缺乏公众参与而难以在制度和手段上回应公众意见;或在强调开放性和自愿性的同时被边缘化而难以发挥其应有的功能。因此,算法"黑箱"的治理手段需要兼顾法律控制性和开放自愿性。强调开放与控制并重的"硬法—软法"范式或可为算法"黑箱"的治理提供一种新的立法框架。

二、"硬法—软法"范式下算法"黑箱"治理的正当性理据

以法律规范的强制力为依据,法律规范可以被分为硬法和软法。硬法是依赖国家强制力保障实施的,具有较强稳定性的规范类型;软法则指效力结构未必完整,无须依靠国家强制力保障实施但能够产生社会实效的,具有较强灵活性的规范类型。⑥ 从理论层面来看,"硬法—软法"范式不仅能够通过建立"软硬并重"的二元法模式完成算法"黑箱"治理的合法性论证,而且能够借助灵活的调整机制,适应不同的算法"黑箱"治理场域,并为政府的算法"黑箱"治理行为提供内在参照标准;从实证层面来看,

① 王雨田主编:《控制论、信息论、系统科学与哲学》,中国人民大学出版社1988年版,第93页。
② Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, Big Data & Society, Vol.3:1, p.1(2016).
③ 参见[美]丹尼尔·贝尔:《后工业社会的来临:对社会预测的一项探索》,高铦、王宏周、魏章玲译,新华出版社1997年版,第480-498页。
④ 参见谭九生、范晓韵:《算法"黑箱"的成因、风险及其治理》,载《湖南科技大学学报(社会科学版)》2020年第6期,第95页。
⑤ 参见[德]乌尔里希·贝克:《风险社会》,何博闻译,译林出版社2004年版,第190-224页。
⑥ 参见罗豪才、宋功德:《认真对待软法——公域软法的一般理论及其中国实践》,载《中国法学》2006年第2期,第3-5页。

借助“硬法—软法”范式治理算法“黑箱”在国外已有先例且取得了不错的实效,可资中国借鉴。

(一)“硬法—软法”范式的理论之维:算法“黑箱”治理的理论依据

算法“黑箱”虽然是一种技术逻辑,但放置于公共治理领域来讨论其规制问题,归根结底其仍然是一个法律问题,首当其冲的应当是承认实体法体系的有效性,充分利用以命令与控制性为主的硬法规范加以治理。但法律发展到了今天,诸多没有国家强制力的软法规范已经在消费者保护、环境保护、技术风险规制等公共治理领域发挥了重要作用,甚至正在重塑国家公域之治的法理基础。① 可以说,软硬法混合在公共治理领域中普遍存在不仅是一个既定事实,更是政府公共部门应当坚持的治理立场和方法。

1.“硬法—软法”范式下算法“黑箱”治理的合法性基础

规范主义模式的合法性基础是宪法至上和法律中心主义,即行政机关的治理行为必须依据法律作出,且不得对宪法中的基本权利造成减损。采用“软硬并重”的二元法治理模式治理算法“黑箱”同样依据这一合法性基础。一方面,无论是硬法规范抑或是软法规范,行政机关的算法“黑箱”治理行为都必须以宪法为根据,不得与宪法相违背;另一方面,为了防止因公权力滥用而对基本权利造成减损,基于硬法的算法“黑箱”治理行为可以通过权力制约和司法审查的方式予以监督,基于软法的算法“黑箱”治理行为可以通过明晰软法制定的公共目的、出台与软法实施相关的规范性文件等方式予以监督。

更为重要的是,在硬法的司法中心主义下,算法“黑箱”治理无可避免地会面临诉讼主体、举证责任等法律形式主义的适用局限。相反,软法在不违背法律中心主义的前提下,能够克服司法中心主义的局限,将算法“黑箱”治理问题引向司法程序以外,并在更大的认同基础上取得合法性基础。这是因为,软法规范虽然不是基于政治性投票形成的,但却在自由市场上经过了公众的充分辩论和调整,是基于自由市场投票而形成的。据此,软法规范的合法性基础可以理解为是国家将专属于自身的公共权力让渡给了自由市场,从而赋予来源于自由市场的软法规范以法律权威。

2.“硬法—软法”范式下算法“黑箱”治理的逻辑机理

目前,“硬法—软法”范式已不仅仅是存在于学理中的范式构想,软法规范在当下的法律实践中已经发挥了重要作用,我们需要将视角聚焦于各类软法规范与硬法规范相互作用,混合治理公共领域的逻辑机理上。一方面,“硬法—软法”范式能够借助灵活的调整机制,适应不同的算法“黑箱”治理场域;另一方面,软法规范能够为政府的算法“黑箱”治理行为提供内在参照标准。

算法“黑箱”的治理需要有公众的参与,但缺乏法律控制的公众参与容易导致规制俘获和有组织的利益压倒,②对此,“硬法—软法”范式能够给出解决方案。硬法通过配置算法“黑箱”所涉各方主体的权利义务和责任关系来展开初次治理,奠定算法“黑箱”规制的合法性基础;软法则通过法律原则、国家政策、行业标准等手段进行二次治理,确保公众参与算法“黑箱”治理以实现公平公正。例如,许多国家都将透明度作为一项原则性规定纳入国家政策或实在法体系,并以此为基础开展算法“黑箱”治理的立法活动。这是因为,一方面,透明度原则的基本精神是明确的,内在蕴含着算法“黑箱”治理的法治立场和价值取向;另一方面,透明度原则作为一项原则性规定,可以根据实际需要,通过立法技术的运用弹性调整法律规则的规制边界,从而适应不同的算法“黑箱”治理场域。

另外,软法规范通过国家性政策、法律原则、法律标准与一系列自我约束的规范性文件能够为政府的算法“黑箱”治理行为提供内在的参照标准。③ 从本质上看,政府的算法“黑箱”治理行为是法律意志借助行政方式得以实现的过程,属于行政法治。行政法治的实现不仅需要司法机关的事后裁断与救济,更需要行政机关的事前防范与引导。④ 因此,政府的算法“黑箱”治理行为不仅应当遵循法律规则的要求,同时应当将法律原则、法律标准作为考虑因素。这里的法律标准既包括风险识别意义上的算法“黑箱”治理标准,也包括执法意义上的政府裁量标准。在行政公共部门的科层式管理体制下,上级行政公共部门的裁量基准能够直接影响到下级行政公

① 参见沈岿:《软硬法混合治理的规范化进路》,载《法学》2021 年第 3 期,第 69-83 页。
② 参见董正爱:《环境风险的规制进路与范式重构——基于硬法与软法的二元构造》,载《现代法学》2023 年第 2 期,第 118 页。
③ 参见[美]罗斯科·庞德:《通过法律的社会控制》,沈宗灵译,商务印书馆 2010 年版,第 27-29 页。
④ 参见张莉:《行政裁量指示的司法控制——法国经验评析》,载《国家行政学院学报》2012 年第 1 期,第 115 页。

共部门的实际决策,且这种行政裁量指示不得违背宪法和法律,这就能够促使宪法与软法规范对政府的算法"黑箱"治理行为进行双重控制,从而防止公权力滥用。此外,作为软法规范的一系列自我约束的规范性文件虽然不具有法律强制力,但可以为政府的算法"黑箱"治理行为提供政策性参照与技术指导。

上述论证表明,"硬法—软法"范式下的算法"黑箱"治理不仅是鼓励公众参与的,而且是法治化的,能够平衡算法研发与应用过程中所涉及的多元利益和价值。其中,硬法规范的初次治理是一种外在规制,其决定了算法"黑箱"治理的裁量边界;软法规范的二次治理是一种内在规制,能够为政府的算法"黑箱"治理行为开放公众参与机制和提供内在参照标准。

(二)"硬法—软法"范式的实证之维:算法"黑箱"治理的域外实践

在新兴科技领域采用"硬法—软法"范式已经是国际通行做法,并且取得了较好的实践效果。国外的相关法律规范多以算法治理或人工智能治理涵括算法"黑箱"治理,且欧盟与美国在此方面走在前列,可以为中国提供相关借鉴。

1.欧盟

欧盟的算法治理虽然以行政机构为主导,以硬法为核心依据,但也通过多方面配置软法以促使利益攸关者参与治理。在算法硬法治理方面,欧盟目前没有单独的算法立法,但基于现代人工智能技术和网络科技的发展,算法无处不在,在算法可能引发问题的个人信息保护、人工智能、消费者权益保护、数字市场等领域,欧盟出台了相关法律法规。这些法律法规涉及到风险评估、数据隐私保护、透明度和可解释性、人权和伦理原则以及监管和合规,包括《通用数据保护条例》(*General Data Protection Regulation*,简称 GDPR)、《电子隐私指令》(*Privacy and Electronic Communications*)、《数字服务法》(*Digital Services Act*)以及《数字市场法》(*Digital Market Act*)等。此外,欧盟还推动制定《人工智能法案》(*Artificial Intelligence Act*),旨在对人工智能技术进行更具体的监管,其中必然涉及算法"黑箱"治理。

在算法软法治理方面,欧盟聚焦于制定人工智能的伦理准则。例如,欧盟委员会 2019 年发布的《可信任人工智能伦理准则》(*Ethics Guidelines for Trustworthy AI*)提出了七个核心原则:人类代理和监督,技术的鲁棒性和安全性,隐私和数据治理,透明度,多样性、非歧视和公平,社会和环境福祉,问责制。① 这些原则旨在为人工智能系统的设计和应用提供指导和规范,确保人工智能系统是可信任、可靠和可持续发展的。类似的软法还有《欧盟人工智能战略》(*Artificial Intelligence for Europe*)、《机器人法规白皮书》(*White Paper on Robotics*)等。

2.美国

在算法领域,美国重视公共部门、私营部门、相关行业等不同主体的力量,通过综合运用法律法规、自治规则、道德规范,形成了多元主体协同、软硬法混合治理的格局。

在算法硬法治理方面,美国采用了"州和地方政府率先立法、联邦政府持续推进"的方式。美国州和地方政府是推动算法治理的先行者,尤其是纽约市,其在 2018 年通过的《算法问责法》开创了美国算法立法治理的先河。纽约市《算法问责法》旨在通过监管政府使用的各类算法以解决算法歧视问题,并推动政府决策算法开源和建立算法问责制。在纽约市率先立法的影响下,2019 年美国国会也引入《算法问责法案》,旨在促进算法开发和应用中的透明度、公平性和问责制。随后,美国国会又在 2019 年《算法问责法案》的基础上提出了 2022 年《算法问责法案》。与 2019 年的版本相比,2022 年的法案不仅更加强调算法的透明度、可解释性、公平性和非歧视性,而且要求算法企业定期报告和审计。②

在算法软法治理上,美国采用了"企业行业自治自律、学术组织/研究机构积极参与"的方式。企业作为算法系统的直接研发和应用主体,社会要求其应当积极履行算法治理责任。为此,微软、谷歌、Facebook、IBM 等科技企业均制定了人工智能伦理原则。例如,2018 年谷歌发布了"对社会有益、避免建立或加剧不公与偏见、保障建立与测试中安全性、对人类负责、建立并体现隐私保护原则、支持并鼓励高标准的技术品格、提供并保障上述原则的可操作性"七项人工智能伦理原则。③ 在行业方面,亚马逊、微

① European Commission, *Ethics Guidelines for Trustworthy AI*, European Commission(8 April 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

② *H.R.6580—Algorithmic Accountability Act of 2022*, Congress.gov(3 February 2022), https://www.congress.gov/bill/117th-congress/house-bill/6580/text.

③ Sundar Pichai, *AI at Google: Our Principles*, Google(7 June 2018), https://blog.google/techno-logy/ai/ai-principles.

软、谷歌、IBM 和 Facebook 联合成立了人工智能合作组织,对算法技术的合规应用提出了行业要求。

学术组织/研究机构也作为第三方参与算法治理。例如,2017 年美国电气和电子工程师协会发布《人工智能设计的伦理准则》(第 2 版),提出了“人权、福祉、问责、透明、慎用”五项基本原则。① 该准则目前已成为国际上最具影响力的人工智能伦理原则的版本之一。同年,美国公共政策委员会计算机协会也发布了《关于算法透明和责任制的声明》(*Statement on Algorithmic Transparency and Accountability*),确定了算法透明和问责制的七项原则,包括“意识、获取和补救机制、问责制、可解释性、可溯源性、可审核性、验证和测试”。②

总的来说,国外基本形成了算法软硬法混合治理的格局(有的国家以硬法为主,有的国家以软法为主)。面对公共部门,强调建立保护制度和强化监管力度,以及时发现算法漏洞和问题并加以解决;面对私营部门,强调研发和应用过程中对技术伦理和后果进行反思,要求其主动承担伦理和社会责任,以实现负责任创新;面对技术人员,引入标准化的算法研发、应用和披露准则,以敦促算法系统的技术开发和使用人员合理研发、应用算法系统并主动披露算法系统的逻辑、目的、潜在影响等;面对社会公众,强调提高算法素养以塑造良好的算法治理环境。这为中国对算法“黑箱”采用“硬法—软法”的二元法治理模式提供了具有参考价值的样本。

三、算法“黑箱”的软法治理

软法规范主要表现为基于公众讨论和行业共识而形成的治理准则或行为准则,如各式各样的“人工智能治理准则”,体现了行业最佳实践,为算法治理提供了改善治理的样板。具体而言,构成算法“黑箱”治理的软法规范应包括增强算法模型的可解释性,向社会披露算法参数以及向社会公开算法源代码。但传统理论框架下,是否遵守任意性规范完全由被规范主体自行决定,这显然无法发挥治理算法“黑箱”的效能。对此,公司治理领域发展出了“遵守或解释”机制,即在任意性规范的基础上结合强制披露义务——不遵守任意性规范时须作出合理的解释性说明,从而在保持任意性规范的灵活性和自主性的同时使其具有一定的约束力。③ 这值得算法治理领域借鉴。

(一)增强算法模型的可解释性

“增强算法模型的可解释性”这一规范在中国已有立法雏形——《互联网信息服务算法推荐管理规定》第 12 条鼓励算法推荐服务提供者综合运用算法设计、优化等方式提高算法模型的透明度,增强算法模型的可解释性。④ 增强算法模型的可解释性的技术路线大体上包括两种,一是直接采用可解释性较强的算法模型,二是采用事后解释的方法增强算法模型的可解释性。需要解释的一点是,增强算法模型的可解释性的技术路线为何不包括采用诸如事先披露数据、算法模型等方法,主要原因是涉及商业秘密保护,下文也会就类似问题展开论述。

就直接采用可解释性较强的算法模型而言,某些算法模型属于自解释模型,其内置解释生成模块,能够对自身的预测结果进行自解释。人类可以从外界较为轻易地观察、检验乃至模拟这些算法模型的运行过程和逻辑。例如,线性回归模型就是一种典型的自解释模型。线性回归模型的求解方法是中国高中数学必修课内容中的“最小二乘法”,具有极高的透明度。⑤ 当然,除了自解释模型外,还有其他可解释性较强的算法模型,如胶囊网络、多粒度级联森林等。

就采用事后解释的方法增强算法模型的可解释性而言,事后解释的技术方法并不是唯一的,可以大体上分为算法模型相关解释和算法模型无关解释两大类。⑥ 顾名思义,算法模型相关解释的对象是特定的算法模型,而算法模型无关解释则不针对某一特定算法模型,可以适用任何算法模型。因此,算法模型无关解释这一技术方法更具灵活性和通用性。对于算法模型无关解释又可进一步分类为规则解

① IEEE Standards Association, *Ethically Aligned Design*, IEEE(12 November 2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

② Association for Computing Machinery US Public Policy Council, *Statement on Algorithmic Transparency and Accountability*, Association for Computing Machinery(12 January 2017), https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

③ 参见安晋城:《算法透明层次论》,载《法学研究》2023 年第 2 期,第 61 页。

④ 《互联网信息服务算法推荐管理规定》第 12 条规定:“鼓励算法推荐服务提供者综合运用内容去重、打散干预等策略,并优化检索、排序、选择、推送、展示等规则的透明度和可解释性,避免对用户产生不良影响,预防和减少争议纠纷。”

⑤ Alejandro Barredo Arrieta & Natalia Díaz-Rodríguez, et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI*, Information Fusion, Vol.58: 82, p.85(2020).

⑥ Vaishak Belle & Ioannis Papantonis, *Principles and Practice of Explainable Machine Learning*, Frontiers in Big Data, Vol.4: 1, p.1(2021).

释、基于特征解释、局部解释和可视化解释。[①] 其中,局部解释(包括局部近似、反事实解释等)是识别算法偏私与歧视的重要技术方案,其有助于披露特定样本与模型输出之间的相互关系,能够有效评价算法自动化决策的公平性。LIME 算法就是一种典型的局部近似的解释方法,其目标是在模型预测的邻域上找到一个可近似代替复杂"黑箱"模型的自解释模型(如线性回归模型),以一种可理解且具有鲁棒性的方式来解释目标模型的预测结果。[②] 古德(Goode)等人就曾采用了 LIME 算法来解释目标模型的预测结果,同时还借助数据可视化中的散点图、评价度量图和热力图来评估 LIME 算法的解释实效。[③]

因此,增强算法模型可解释性的路径是多元化的,包括采用自解释模型、采用可解释性较强的算法模型、提供有效的事后解释(局部近似、反事实解释等)方法。

基于技术中立和技术自由发展原则,"增强算法模型的可解释性"只能设置为软法规范,由算法开发者/运营者依据其实际情况自行作出选择,同时,要求未遵守该规范的算法开发者/运营者作出合理解释说明。

(二)向社会公开算法源代码

治理算法"黑箱",提高算法透明度的最激进的解决方案莫过于向社会公开算法源代码(也被称为"算法开源"[④])。在网络世界里,代码/架构的设计者能够建立网络空间的默认规则,决定着网民的自由程度和范围,相当于网络空间的立法者。[⑤] 因此,对于是否应当规制代码,一直存在争议。在早期互联网发展中,互联网传输遵循的是端对端交互的规律,没有规制代码的空间,且那些为搭建互联网而编写了代码的作者出于维护自身利益的目的,也会对政府公共部门的代码规制行为加以抵制。[⑥] 但当前代码的编写已经日益商业化并集中于少数私人资本控制的大型互联网公司中,这就意味着这些私人资本控制的大型互联网公司借助代码成为了网络空间的立法者。此种情形下,若再不对代码加以规制,网络空间就会产生商业利益驱逐公共利益的隐患。在中国已经步入算法社会的当下,政府的职责是要确保私人资本控制的商业机构在设计代码/架构时能够纳入公共利益。尽管政府公共部门无法直接规制代码,但可以通过规制私人技术公司来实现间接规制代码的目的。另外,算法源代码的公布虽然只是暴露了算法运行所采用的机器学习方法,尚无法解释算法决策的过程,但相关专业技术人员可以通过检验算法源代码,判断算法运行和决策过程中是否存在算法歧视、算法决策缺陷等问题,从而达致"鱼缸透明"的效果。[⑦] 这些现实情况都为算法开源奠定了基础。

在互联网开放共享的精神下,一些大型互联网公司已经在网络世界的开源社区中公开了相关项目以及算法组件的源代码。[⑧] 算法源代码的公开不仅能够增强普通公众与专业技术人员对算法系统的信任,而且有利于算法企业借助社会上大量专业人士和爱好者对源代码的优化,从而迅速在市场竞争中占据优势。比如 Linux 操作系统的代码就是开源的,公众不仅可以公开访问、查看、复制源代码,甚至可以通过特定的电子邮件列表参与到对源代码的修改和重新分发环节中。集公众之长对 Linux 操作系统的源代码进行迭代更新也成为了该操作系统长期保持安全稳定,进而占据服务器市场超过 80%份额的重要原因之一。长期以往,开源社区、开源项目和开源代码逐渐会形成一个竞争性的算法代码市场,而自由竞争的算法代码市场也会"倒逼"私人技术公司公开算法源代码。因此,国家不必强制算法开源,借助自由市场的力量亦可达到算法开源的效果。

更为重要的是,强制算法开源会面临商业利益和公共利益的冲突问题。早在 2016 年,纽约市就曾经欲借助立法强制政府公共部门披露算法源代码,但相关立法议案一经提出即遭致多方反对:企业认

---

① 参见靳庆文、朝乐门、孟刚:《AI 治理中的算法解释及其实现方法研究》,载《情报资料工作》2022 年第 5 期,第 16-18 页。

② Marco T. Ribeiro, Sameer Singh & Carlos Guestrin, *Why Should I Trust You? Explaining the Predictions of Any Classifier*, in Balaji Krishnapuram & Mohak Shah, et al. eds., Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p.1135-1144.

③ Katherine Goode & Heike Hofmann, *Visual Diagnostics of an Explainer Model: Tools for the Assessment of LIME Explanations*, Statistical Analysis and Data Mining: The ASA Data Science Journal, Vol.14: 185, p.185-200(2021).

④ "算法开源"意味着任何人都可以公开访问、查看、修改和重新分发算法源代码。

⑤ 参见[美]劳伦斯·莱斯格:《代码 2.0:网络空间中的法律》(第 2 版),李旭、沈伟伟译,清华大学出版社 2018 年版,第 152 页。

⑥ 参见汪庆华:《算法透明的多重维度和算法问责》,载《比较法研究》2020 年第 6 期,第 170 页。

⑦ Joshua A. Kroll & Joanna Huey, et al., *Accountable Algorithms*, University of Pennsylvania Law Review, Vol.165: 633, p.638(2017).

⑧ 参见安晋城:《算法透明层次论》,载《法学研究》2023 年第 2 期,第 62-63 页。

为源代码是算法技术的核心,披露算法源代码会侵犯企业的核心商业秘密;一些网络安全专家也指出算法源代码的公开很有可能会为不法分子侵害公共安全提供技术便利,甚至可能衍生“戏耍”算法的风险。在这些反对的干预下,强制政府公共部门披露算法源代码的提议并未出现在纽约市最后通过的算法立法要求中。这足以反映出算法开源内在潜藏的商业利益与公共利益之间的紧张关系。因此,出于调和商业利益和公共利益之间的冲突的目的,国家不宜强制要求算法企业向社会公开算法源代码。① 当然,这并不与政府公共部门规制算法源代码相悖,盖因法律规范除了包括具有强制力的硬法规范外,还包括具有灵活性的软法规范。

因此,中国应在现有的开源实践、公众讨论和行业共识基础上制定算法开源准则,②同时要求未遵守算法开源准则的算法开发者/运营者作出合理的解释说明。

四、算法“黑箱”的硬法治理

除软法规范外,算法“黑箱”的治理必然离不开硬法规范的支持,否则其治理功能很有可能会被架空,甚至沦落为“倡导性条款”,无法实现透视算法“黑箱”的目的。对于算法“黑箱”的硬法控制,可以从算法权利和算法问责两方面展开论述。算法权利是对算法“黑箱”的前端法律控制,其主要是借助权利对抗中的制约以及权利平衡中的合作,形成对算法开发者/运营者的约束和对算法相对人自身行为模式的调适,从而改变算法关系中的力量对比和利益结构,达到提高算法透明度,治理算法“黑箱”的效果。算法问责是对算法“黑箱”的后端法律控制,其目的是建立适用于算法研发与应用过程中各方主体的责任体系,以便在算法运行过程中出现精确性与公平性偏差时,政府公共部门得以依据相关信息及时确定责任归属。

(一)算法权利

当前法学界对于算法权利的研究与尝试颇为丰富,算法解释权、算法自动化决策拒绝权、算法排他权、人工接管权等权利类型不断丰富和完善。③ 在算法权利的配置中,与算法“黑箱”治理最为密切的是算法解释权和算法自动化决策拒绝权。

关于算法解释权。算法解释权的核心功能是保障算法相对人能够针对算法的设计目的、功能逻辑、决策结果来源等关键信息获悉算法开发者/运营者解释的权利,从而提高算法的透明度。④ 利用算法解释权可以在算法相对人与算法开发者/运营者之间构造对抗型法律关系,促使算法相对人能够借助算法解释权尽可能发现和排除因算法“黑箱”引起的算法偏私与歧视。同时,算法解释权的相关立法也可以“倒逼”私人技术公司探索透明度更高的算法技术。

对于算法解释权的立法与实践,最为关键的一点是明确算法解释的技术标准,即算法开发者/运营者需要向算法相对人解释什么内容和解释到什么程度。算法解释原则上不要求向算法相对人充分解释算法从输入到输出的全过程,而是侧重于以可理解的方式向算法相对人解释决策结果以及决策时可能影响相对人合法权益的相关因素。⑤ 这是因为,有效的机器学习算法模型会对其内在决策模型和代码内容进行实时动态调整或迭代更新,并通过自主学习实现持续复杂化和深度拟人化,即便是算法开发者/运营者也难以详细解释算法运行和决策的全过程。⑥ 因此,算法解释的技术标准可以总结为:算法开发者/运营者应通过可理解的解释方式,达到打消算法相对人对决策结果、决策时可能影响其合法权益的相关因素的疑虑之效果。

关于算法自动化决策拒绝权。算法自动化决策拒绝权能够保障算法相对人有摆脱算法“黑箱”影响的选择自由,也能够对算法开发者/运营者提高算法透明度给予制度激励。前者是需求侧的影响,本质上是算法相对人对缺乏透明度或透明度不高的算法自动化决策“用脚投票”⑦,以规避因算法“黑箱”可能给其合法权益带来的潜在不利影响。后者是供给侧的考量,旨在促使算法开发者/运营者有足够的动力提高算法的透明度,最大程度减少算法“黑箱”的影响。盖因算法开发者/运营者要避免算法相对人摆脱和拒绝算法自动化决策,维持其在算法市场上的占有份额,就只能尽可能提高算法透明度并以

① 参见汪庆华:《算法透明的多重维度和算法问责》,载《比较法研究》2020 年第 6 期,第 170-171 页。
② 参见安晋城:《算法透明层次论》,载《法学研究》2023 年第 2 期,第 63 页。
③ 参见袁康:《可信算法的法律规制》,载《东方法学》2021 年第 3 期,第 17 页。
④ 参见许可、朱悦:《算法解释权:科技与法律的双重视角》,载《苏州大学学报(哲学社会科学版)》2020 年第 2 期,第 67-68 页。
⑤ 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 70 页。
⑥ Joshua A. Kroll & Joanna Huey, et al., *Accountable Algorithms*, University of Pennsylvania Law Review, Vol.165: 633, p.638-639(2017).
⑦ “用脚投票”是一个通俗的表述,用来形容人们对于某个产品、服务、政策或观点等的抵触、放弃或反对。

可理解的方式向公众解释算法的运行和决策过程。

鉴于算法自动化决策拒绝权针对的是“算法决策”,而非“算法决策的结果”,立法应当从事前和事中两个阶段为个体提供履行算法自动化决策拒绝权的保障措施。在事前阶段,事前知晓是个体对抗算法自动化决策的源头途径。《中华人民共和国个人信息保护法》(简称《个人信息保护法》)第14条、第17条、第44条赋予了信息主体知情权和决定权,其中第14条是规定信息主体知情权的原则性条款;第17条则在第14条的基础上进一步细化了个人信息处理者应当告知信息主体的具体事项范围,包括处理的目的、方式、个人信息种类和保存期限,但立法所明确的告知的具体事项范围尚无法完全满足信息主体对算法自动化决策的知情需求,且这些事项范围仅为常规性事项,无法对标对个人权益有重大影响的算法自动化决策;第44条仅简略规定信息主体的知情权和决定权,明确信息主体有权限制、拒绝他人对其个人信息的处理,但尚未阐明如何知情和决定,如何限制和拒绝。除前述条文外,《个人信息保护法》第48条还赋予信息主体获取解释说明的权利,但同样地,如何获取解释说明仍然语焉不详。故后续立法应当为信息主体的知情权和决定权配套相关细则,从事前知晓的角度为个体拒绝算法自动化决策提供源头保护。

在事中阶段,应借助《个人信息保护法》第44条至第47条所形成的包括删除限制、查阅复制、更正补充等权利在内的权利束,为个体提供拒绝算法自动化决策的多样化路径。此外,《个人信息保护法》第24条第2款还要求进行算法自动化决策的行为者在信息推送、商业营销的场景下,应当向个体提供不泄露个人信息的选项设置以及便捷的拒绝方式。① 上述权利一定程度上在事中阶段为个体对自动化决策施加人为限制因素提供了基础。但中国对删除限制、查阅复制、更正补充等权利的立法规定在条文设计上仍略显粗糙,还可再行精细化处理。同时,特定场景下的为个体提供不泄露个人信息的选项设置以及便捷的拒绝方式的立法设计尚不具有普适性,可将此项规定向普适性条款的方向作进一步完善。

(二)算法问责

实践中,不少私人资本通过算法以“作为”的形式实施了诸多违法行为。例如,今日头条旗下的“内涵段子”应用程序及公众号的推荐算法就坚持“只要价值,不要价值观”,肆意传播充斥着色情、暴力、低俗的内容。但传统的追责逻辑遵循的是“主体—行为—责任”,而算法运行所致危害后果的根源通常是多方面的,例如数据来源和质量、代码变量选择、权重设定、架构设计等的偏差均有可能造成损害结果。对此,政府公共部门在启动事后追责机制时,囿于算法“黑箱”的影响,将会难以确定责任来源的根源,更无从考究算法开发者/运营者是否具有过错。由此,国内外立法实践②和理论研究③纷纷提出应启动算法问责。美国的《算法问责法案》、欧盟的《算法问责及透明度监管框架》等都是推动算法问责相关立法的重要体现。

算法问责的责任主体应为算法开发者/运营者。算法运行造成的危害后果也许并非是算法开发者/运营者的主观过错所致,而是多方作用的结果,甚至可能算法开发者/运营者也无法预测这些危害后果。但这不足以免除算法开发者/运营者的算法责任,盖因无论算法如何基于深度学习自主演变,算法开发者/运营者都对算法演变负有一定义务,包括在算法系统中嵌入审计日志,对算法系统进行备案、评估、测试、监督,甚至在算法运行过程中增加权限控制等,以尽可能避免损害结果发生。④

一方面,算法开发者/运营者的主观意图对算法决策结果存在着根源性影响。算法开发者/运营者或多或少会利用代码内嵌规则在算法开发和运营过程中嵌入自身的价值观念或主观意图。在“*Search King v. Google* 案”⑤中,Google就毫不避讳地承认其是有意篡改网页排名算法。另外,即便算法的运行规则与方式导致输出结果可能无法为开发者/运营者所控制,但算法通过深度学习后输出结果的整体技术路线依然是遵循目标导向原则,即以无限接近算法开发者/运营者所设定的输出目标为指引。

① 《个人信息保护法》第24条第2款规定:“通过自动化决策方式向个人进行信息推送、商业营销,应当同时提供不针对其个人特征的选项,或者向个人提供便捷的拒绝方式。”

② 国内外关于算法问责制的立法实践包括2019年欧盟的《算法问责及透明度监管框架》、2022年美国的《算法问责法案》等。

③ 国内外关于算法问责制的理论研究参见 Joshua A. Kroll & Joanna Huey, et al., *Accountable Algorithms*, University of Pennsylvania Law Review, Vol.165: 633, p.638-705(2017);汪庆华:《算法透明的多重维度和算法问责》,载《比较法研究》2020年第6期,第163-173页;张凌寒:《网络平台监管的算法问责制构建》,载《东方法学》2021年第3期,第22-40页。

④ Luciano Floridi, *Distributed Morality in an Information Society*, Science and Engineering Ethics, Vol.19: 727, p.728(2013).

⑤ *Search King, Inc. v. Google Technology, Inc.*, Case No.CIV-02-1457-M(W.D.Okla.May 27, 2003).

另一方面,算法开发者/运营者对算法决策结果负有注意义务。算法开发者虽然无法完全预测和控制算法输出结果,但其在开发过程中应当能够预设合理的数据集选取和质量变量选择、权重设定、架构设计等因素的偏差范围,以及应当具备应对不良算法决策的必要反应能力,从而最大限度确保算法能够应对潜在的风险。算法运营者虽然未参与算法的开发设计,但依据《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》)等法律法规的要求,算法运营者负有采取必要手段(备案、审查、复核、验证等)确保算法决策结果合法合规的义务。例如,在“蚂蚁金服诉企查查案”①中,一审法院与二审法院均认为:大数据企业有确保其收集、发布的数据质量的义务,尤其是对于重大负面敏感数据,应当通过改进算法技术、数据复核、交叉验证等手段,提高数据推送质量,避免不当的信息推送行为。

当然,有效的算法问责还有赖于清晰的主观过错认定,需要借助算法备案、影响评估与合规审计,从算法开发/运营的目的、风险评估以及风险控制能力三方面固定算法问责点。

关于算法开发/运营的目的的问责,算法开发者/运营者对自己开发/运营的算法负有注意义务,应对算法开发/运营的目的及其应用作必要性评估并备案。尤其是当一个算法系统有多个价值目标时,鉴于算法系统缺乏价值判断能力,无法自主协商有冲突的价值目标,算法开发者/运营者在开发/运营算法系统时需要明确这些价值目标的顺位,并将这些不同价值目标的优先级进行评估与备案。② 自动驾驶汽车面临的“电车难题”反映的就是算法价值目标的顺位与取舍问题。

关于算法开发/运营的风险评估的问责,其关键是要确定应实现何种程度的风险评估。算法开发/运营的风险评估最起码应包括算法风险的来源,性质,级别,不同风险级别的算法对公民权利、社会公共利益可能产生的影响。当然,域外一些法律文件也提出了更广泛意义上的,包括对人权、隐私和数据保护等方面的算法影响评估。③ 因此,应在基础性评估内容的基础上,根据算法的功能和场景应用进一步确定风险评估的要求。

关于算法开发/运营的风险控制能力的问责,合规的算法开发/运营的风险控制能力要求算法开发者/运营者应当在风险评估后确定相应的风险防控方案和措施并进行备案,其中应当涉及算法系统的技术逻辑、风险级别与影响评估、风险处理等措施的置备及相关信息的留存。此外,无论算法开发/运营的风险评估为何,算法开发者/运营者均应当事前设置算法的“紧急制停预案与措施”——当算法运行过程中面临公民合法人身权益或重大财产权益损害的风险时,算法开发者/运营者能够紧急中断算法运行与输出的预案与措施。

五、结语:基于技术标准的算法“黑箱”软硬法治理谱系

诚然,强调开放与控制并重的“硬法—软法”范式能够为算法“黑箱”的治理提供一种新的立法框架,但在“硬法—软法”范式下除了需要明确算法“黑箱”治理的硬法规范和软法规范外,还需要在两种规范中建立双向沟通机制,确保两种规范的衔接与协调。此外,硬法是由国家强制力保障实施,而软法是不具有国家强制力支持的法律规范,如何使软法能够在算法“黑箱”治理中得到有效实施,也需要进一步研究。为此,需要重点考虑以下几个问题。

第一,将技术标准作为软法与硬法的沟通机制。算法本质上是一种技术工具,其“黑箱”治理在软法规范上应包括增强算法模型的可解释性和向社会公开源代码,那么,无法回避的是技术标准问题,包括达到何种标准才满足增强算法模型可解释性之要求,向社会公开源代码的具体技术要求是什么等问题。技术标准作为一种典型的软法,由国家标准、行业标准、企业标准等构成,是技术与法治的耦合,是科学技术与价值判断相融合的合理化过程,可以作为规制与自治之间的衔接点和缓冲区。④ 国家部委及地方政府等监管机构牵头制定的技术标准可被视为硬法的触手,用以约束或引导生产生活;另有相当数量的技术标准源于对最佳实践的总结,具有自下

① 参见浙江省杭州市中级人民法院(2020)浙 01 民终 4847 号民事判决书。

② James Mcgrath & Ankur Gupta, *Writing a Moral Code: Algorithms for Ethical Reasoning by Humans and Machines*, Religions, Vol.9: 240, p.240-259(2018).

③ 纽约大学 AI Now 研究所、2022 年美国的《算法问责法案》、2019 年欧盟的《关于先进数字技术的人权影响框架》提出的算法影响评估框架中包括评估对人权的影响;欧盟 GDPR 蕴含了需评估算法自动化决策对隐私的影响,同时要求启动“数据保护影响评估”。相关法律文件还有欧盟的《可信赖的人工智能伦理准则》《算法问责及透明度监管框架》、联合国教科文组织的《机器人伦理报告》等。

④ 参见陈伟:《作为规范的技术标准及其与法律的关系》,载《法学研究》2022 年第 5 期,第 86-87 页。

而上的自治色彩。[①] 因此,技术标准适合作为沟通机制,用以建立算法“黑箱”的软硬法治理体系。

从比较法视野看,国外对于借助技术标准将软法嵌入硬法并形成利益导向机制已具备规模且形成体系化的经验。国际标准化组织与国际电工委员会联合发布的 ISO/IEC 27701 标准就是可资借鉴的典型示例。制定 ISO/IEC 27701 标准的目的是为了通过附加要求增强现有的信息安全管理系统,建立、实施、维护和持续改进隐私信息管理系统,以便遵守 GDPR 并满足其他数据隐私要求。ISO/IEC 27701 标准的附录 D 主要阐述其与 GDPR 的映射关系,表明相关主体遵守标准中的要求和控制措施与其履行 GDPR 的相关性,甚至明确表示遵守标准中的单个隐私控制点可以满足 GDPR 中的多项要求。虽然 ISO/IEC 27701 标准附录 D 所阐述的与 GDPR 的映射关系仅具有指引作用,但这些映射关系的制定是由 GDPR 的立法委员会参与的,具有一定的权威性。由此可见,GDPR 与 ISO/IEC 27701 标准之间形成了成文法与技术软法的嵌入互动机制,链接了“技术”与“法律”两个话语体系。因此,中国可借鉴 GDPR 与 ISO/IEC 27701 标准的经验,借助技术标准链接算法“黑箱”治理的软法与硬法,建立符合中国国情的算法“黑箱”软硬法治理体系。

第二,在立法上塑造技术标准生成的硬法环境。欲在算法“黑箱”治理中充分发挥软法的独特功能,中国需形成完整的算法治理法律规范体系,否则没有明确硬法映射的软法也将失去其规范优势。当前中国针对算法治理的立法散见于《中华人民共和国民法典》(简称《民法典》)、《中华人民共和国电子商务法》《个人信息保护法》《网络安全法》《数据安全法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等法律、行政法规和部门规章中。《浙江省数字经济促进条例》《深圳经济特区数据条例》《深圳经济特区人工智能产业促进条例》等地方性法规也有相关规定。这些法律法规都可以直接或间接治理算法“黑箱”。

然而,现行立法尚无法为算法“黑箱”治理的软法规范提供充分的硬法映射:《民法典》的规定具有概括性,能够奠定算法规制的基本框架,但主要通过赋权的方式从数据、个人信息保护的角度间接规制算法,义务要素结构有所缺乏;《数据安全法》和《网络安全法》虽然具备一定的义务要素结构,但二者主要聚焦宏观的网络安全和数据安全,与算法“黑箱”治理并非完全对应关系,难以直接作为硬法来映射;《个人信息保护法》具备一定的义务要素结构,且明确对算法自动化决策加以规制,并形成了算法影响评估、审计和备案制度的雏形,但仍然缺乏配套措施。即便是与算法“黑箱”治理最密切相关的《互联网信息服务算法推荐管理规定》,其规制对象也限于互联网信息服务算法推荐技术,且主要围绕算法安全主体责任、用户权利以及算法分类分级安全管理制度,并未有增强算法模型可解释性以及向社会公开源代码的相关规范,同样难以直接作为硬法来映射。因此,需要完善中国算法“黑箱”治理的法律规范体系,从立法的角度塑造技术标准生成的硬法环境。

第三,在司法上明确技术标准嵌入的可能方式。技术标准可通过法律适用与事实认定两方面嵌入司法。在法律适用层面,技术标准可在法律存在空白时起到间接适用功能,在法律不确定时起到解释补充功能;在事实认定层面,技术标准可作为书证。技术标准嵌入司法在中国已有先例。例如,在“葛明君诉中国银行股份有限公司成都科华街支行案”[②]中,被告中国银行科华街支行证明中国银行手机客户端已经过专门的认证机构进行安全性认证且符合监管规定及行业标准,并以此作为已履行安全保障义务的充足证据,避免了责任承担。在“北京金星鸿业电梯有限公司诉北京市朝阳区质量技术监督局案”[③]中,法院根据《JG135—2000 杂物电梯》和《TSG T5001—2009 电梯使用管理与维护保养规则》辅助认定电梯维护保养单位的维保义务。总的来说,在事实认定层面,技术标准可以作为一种书面证据,用以证明算法开发者/运营者在事实层面具备算法开发/运营的资质且体系化地履行了现有法律规定的注意义务(如遵守 ISO/IEC 27701 一定程度上即遵守了 GDPR)。在法律适用层面,技术标准可以在法律规定的义务内容、过错认定条件较模糊导致自由裁量空间过大时,作为规则的补充释明,例如,可以根据技术标准的规定判断义务履行主体的过错程度,酌定赔偿数额等。

① 参见衣俊霖:《数字孪生时代的法律与问责——通过技术标准透视算法黑箱》,载《东方法学》2021 年第 4 期,第 88 页。
② 参见四川省成都市中级人民法院(2021)川 01 民终字 18180 号民事判决书。
③ 参见北京市第二中级人民法院(2011)二中行终字第 937 号行政判决书。