算法治理的黑箱及规制：基于治理界面的视角*

刘　建　吴理财

摘要：治理界面不仅是算法治理的情境空间，也为理解算法治理黑箱提供了一个重要视角。算法与复杂性治理情境相结合以后，生成了一种混沌的治理界面，形塑了技术、认知与关系为一体的黑箱结构。在这一结构中，高密度的数据资源为算法治理黑箱提供持续运转的能量，敏捷式界面装置实现了黑箱技术的迭代优化，反思性的界面调节则促成了黑箱的自我庇护，三者共同构成了算法治理黑箱实现界面控制的内在机制。然而，技治主义的界面设计、工具理性的界面规范和内外有别的界面运行方式，也在不断固化算法治理黑箱的界面，使其陷入“进退两难”的困境。为此，有必要将界面治理纳入算法治理议题之中，通过界面修复、净化及共生机制的建构对算法治理黑箱进行规制，探索算法治理信任体系的重建路径。

关键词：算法治理　算法黑箱　黑箱规制　治理界面

中图分类号：D630;TP301.6　　**文献标识码：**A　　**文章编号：**1004-0730(2025)11-0034-12

DOI:10.19624/j.cnki.cn42-1005/c.2025.11.002

一、引言

以人工智能算法为代表的信息技术正在深刻重塑着日常生活，人类社会由此进入算法主导的时代。在这一对社会基础设施进行全面重构的过程中，由于算法本身的不透明性、复杂性及潜在的歧视性等因素的相互交织，逐渐催生了算法黑箱问题[1]。算法在输入、输出及运行过程中的不透明性与不可解释性，使公众与社会对“算法黑箱”普遍产生了深刻的焦虑[2]。需要注意的是，算法自其诞生以来，就作为一种治理工具及治理场域而存在，随着信息技术的迭代式发展，算法治理已经演化为一种新型的治理范式。在政府决策、经济运行、社会监管与文化传播等领域，算法均提供了全新的治理机制。然而，伴随着算法在公共领域的深度嵌入，原先的算法黑箱问题也进一步延伸并转化为算法治理黑箱。

如何剖析算法治理黑箱的内在结构并在此基础上对其实现有效规制，是算法治理研究的重要议题。当前，学界对此主要呈现为“技术-制度”的双重解释视角，将其视为技术与规则相互交织的产物[3]。在技术治理视野下，学者们指出，由于算法无法真正保持技术中立原则[4]，其存在

的对公众而言难以理解的内在逻辑或曰“算法隐层”[5]便成为一个突出的治理难题，容易引起个体主体性异化及传统决策治理框架失灵的风险[6]。在治理规则视野下，学者们认为，由于算法在公共治理场域的“规则隔音”，算法治理陷入了信息不透明性和不可解释性的困境[7]。为有效规制算法治理黑箱，应在算法治理场域嵌入协商民主理念[8]，确立算法治理逆向评测和监管体系[9]，以算法问责制来形塑“负责任的算法”[10]。

从算法黑箱到算法治理黑箱这一研究视野的拓展，不仅反映了公众对算法治理风险的普遍关切，也为理解算法治理黑箱拓宽了理论资源。要揭示算法治理黑箱的形成机理，不仅需要从“技术-制度”维度进行理解，更应从其形成的内外部环境展开思考。为此，有必要从算法治理黑箱的要素构成出发，揭示其整体的系统机制，进而阐释其生成逻辑。算法治理黑箱不仅是一个技术问题，更是一个治理问题，其本质是算法系统与治理系统在特定界面耦合的产物。这里的“界面”可被理解为由内部环境与外部环境共同塑造的“人工物”：内部环境是界面的组织模式及运转方式，外部环境则是界面运行的外生空间[11]。界面构建由此成为治理的前提，内部结构、界面、目标与环境则是治理界面理论的四个基本要素[12]。从这一视角出发，算法治理黑箱亦可被视为一种治理界面黑箱，是算法黑箱拓展至公共治理场域后，与治理情境耦合形成的产物。基于此，本文将在治理界面的理论视域下，系统阐释算法治理黑箱的结构体系与生成机理，并进一步探索其可能的规制路径。

二、界面的混沌：算法治理黑箱的存在方式

算法治理黑箱的存在方式与实践面向，表现为其内部结构相对稳定而外部主体却无法清晰窥探其运作逻辑，从而呈现出一种界面混沌的状态。当算法将数据装进技术代码的“匣子”之后，其在实践中便展现出符号代码、信息控制与权力表达的多重黑箱维度[13]。随着算法黑箱与复杂治理情境在界面相遇，算法黑箱的类型结构也随之转化为算法治理黑箱的实践面向，生成了一种将技术性、观念性及关系性融为一体的复合治理界面，由此形塑了算法治理黑箱的存在方式。

（一）技术性黑箱：作为实体空间存在的界面

算法的天然黑箱是算法治理黑箱生成的内在前提，而复杂的治理场景则构成了其外在生成变量。在算法系统内部，物理实体、数据资源及算法模型等要素构成了一个闭环的物理通道，数据的输入、处理和输出均在其中完成，技术代码由此转化为能动性的治理程序与机制。具体而言，技术代码在对冗余数据进行剔除的过程中，将目标数据以特定格式进行转换，进而达到数据治理集约化及界面抽象化的目标。换言之，算法将程序性知识封装进技术代码内部，在数据输入与输出之间构建了一个具有模糊性的“隐层”；该“隐层”的界面设计及工作原理对于设计者而言相对清晰，但对于公众而言则构成一个天然的、难以理解的技术黑箱。在此基础上，算法的技术代码与公共治理技术在界面的耦合，进一步生成了一种技术性的算法治理黑箱。

在这一黑箱的要素构成中，技术代码及承载代码的算法平台构成其“黑匣子”；算法治理将治理行为视为各种信息流交汇的结果，通过对治理内容进行标准化与格式化处理，海量数据被转化成一种黑箱化的底层代码和计算单元[14]。在对数据进行收集、整理及分类的过程中，代码技术与治理技术被同步嵌入黑箱内部，进而在技术层面确立了算法治理黑箱的物理空间。人工智能尤其是机器学习技术在政府、企业及非营利部门的广泛应用，显著强化了算法治理黑箱的技术特

征；但智能算法在优化治理决策的同时，也塑造了一种技术构造更为封闭的界面。一言之，技术代码为算法治理黑箱提供了“技术集装器”，使算法的天然技术黑箱传导至算法治理场域；复杂性治理情境与算法技术黑箱的叠合，更加掩盖了治理决策过程，甚至导致决策者乃至算法设计者都无法全部厘清决策依据，从而加剧了算法治理黑箱化运转的倾向。

（二）观念性黑箱：作为认知模糊存在的界面

“黑箱”一词不仅蕴含着技术代码的混沌特性，更体现出公众对算法治理界面认知的混沌形态。从日常生活观念的视角来看，算法治理界面不仅包括系统的自我编码与解码，还体现了人机交互背景下界面主体的认知状态。一方面，算法黑箱具有类似人类大脑的认知能力，只有通过输入和输出才能了解它的工作原理[15]。另一方面，智能算法在算法治理界面中嵌入了类神经元计算单元系统，这些系统模拟了人体神经的连接与信号处理功能，能够将信息数据转化为可计算的数值形式。然而，算法设计、开发及应用分散于不同的研发主体，数据治理也涉及多元参与方，多种算法的叠加及多个隐含层的存在加剧了算法的不可解释性风险，由此在输入与输出之间形成了无法解释的“黑洞”[16]。算法通过这种输入与输出之间的“屏蔽”界面实现了信息控制，由此强化了算法治理的认知隔阂，从而形塑了一种社会认知层面的黑箱。

算法治理黑箱的社会存在方式，既源于算法治理本身的技术复杂性与认知遮蔽特性，也与公众面对“陌生性技术”时产生的不确定性恐惧密切相关。在智能算法的助推下，机器学习不断向自动化及拟人化方向发展，公共治理领域中自动化决策的特征日益显著。尤其值得注意的是，深度学习技术驱动了计算机代码的自动化运行，使算法在一些领域具备了类似甚至超越人类的“认知智能”；与此同时，人类对算法的“输入”“输出”环节的感知及把控能力却不断减弱，进一步加剧了公众对算法治理黑箱的认知焦虑。信息的不透明与机制的难以解释，不仅在认知与决策层面凸显了算法治理的黑箱特性，更塑造了一种以认知模糊为核心特征的、混沌的观念性界面。

（三）关系性黑箱：作为复合权力存在的界面

算法并非在社会真空中运行，其本身就是社会关系抽象化的表征，亦即在治理界面中建构了一种对社会关系的“抽离机制”。在此基础之上，在形塑实体性和认知性界面的同时，算法还构造出一种关系性界面，进而为算法时代的权力关系实践提供了载体。算法作为解决特定问题的代码集合，无论是数据筛选标准的制定，还是数据选择的偏好乃至输出结果的预期，背后都反映出权力实践的逻辑。具体到社会治理领域，算法将各项国家治理事务转换为可标准化、程式化处理的治理因子，以可计算化的逻辑重构了日常生活的关系图谱。换言之，算法治理将计算逻辑嵌入治理决策系统，从而决定了治理场域中哪些行为可被采纳、限制或推广，使算法在国家治理体系中具备了规范与分配资源的权力。由此，算法权力可被视为一种控制力，体现为技术平台的研发者和控制者利用自身技术优势生成的一种技术影响力[17]。随着算法在国家治理领域的深度嵌入，技术代码权力与国家权力进一步相互交织，使得技术性权力与关系性权力均被纳入算法权力谱系之中，进而形成了一种更为复杂的权力结构。这种结构已经超越技术或者认知单一层面的混沌性，演变为一种复杂的社会关系网络，一种由异质行动者关联互动形成的“关系性存在”[18]。

算法系统通过对各种数据要素进行结构化再生产，不仅改变了传统权力关系的可见性，更重塑了其内在的运作机制，实现了社会-技术关系的重组，推进了算法治理时代权力实践范式的变革。算法以一种“隐身术”的形式弱化了权力

的强制性面向，在信息内容与服务的个性化配置过程中建立起一种软性支配关系。由此可见，算法不仅是一个确定性的技术黑盒，更是一个庞大的、网络化的、与社会现实高频互动的系统；其在嵌入社会结构的过程中，进一步加剧了自身的混沌性[19]。特别是当多种算法进行优化组合时，其所需要处理的数据及数据之间的变量关系难以被直观的方式呈现，治理情境的复杂性则更为加剧了权力关系的复杂性，算法本身难以理解的认知特性又进一步强化了算法权力的赋魅效应，进而导致了算法治理中权力关系的黑箱化运作。算法治理界面作为治理系统内外部要素交流耦合的纽带，也在这一过程中成为承载这种复合型混沌权力关系的重要场域。

三、界面的控制：算法治理黑箱的自我维护

在算法治理黑箱的内部结构中，生成了集数据输入、处理及输出于一体的治理链条，从而建立起一套界面控制机制。具体而言，高密度的界面要素持续为算法治理黑箱提供能量维持，敏捷式的界面装置实现其技术优化，反思性的界面结构则推进其自我调适，三者共同为算法治理黑箱的有效运转奠定了自我维护体系。

（一）高密度的界面要素：算法治理黑箱的能量维持

在社会系统的运转中，算法治理黑箱要实现界面的自我控制，必须在内部建立一套持续获取能量的动力机制。信息数据作为算法治理的“生命源泉”，经标准化处理器加工后，由平台、流程与算法协同聚合为高密度要素，由此构成了基础的数据能量供给链条。为进一步强化这一能量供给，黑箱借助传感器强大的数据收集与流通能力，对更广泛的数据进行整合。这使得黑箱的整体运行模式，从基础的数据聚合演进为一种高效的“数据共生”状态；它能将不同轨道的数据动态汇聚并融合于统一的算法治理界面之中，从而获得规模更大、更稳定的能量支持。

高密度的界面要素不仅体现在数据规模上，也体现在算法治理平台强大的数据治理能力中。具言之，算法治理平台通过将各类数据按照特定标准整合为相互联结的治理网络，为算法治理黑箱的建构及运转提供了稳定的界面环境。平台还保障了数据信息在不同治理节点之间的流通，使得原本杂乱无章、质量参差不齐、分布于多个平台和服务器的大数据，通过“结构洞”的方式实现了重塑[20]。然而，作为算法治理黑箱的“发动机”，数据治理平台又是以一种社会难以理解的方式在进行运转，为算法黑箱提供了“自我保护”界面。在此基础上，算法治理平台将治理知识谱系、治理流程、时空维度有机整合，从而把多元化的社会情境纳入算法治理系统之中，在算法治理黑箱内部构建起了一种能量稳定供应的界面控制机制。从结构上看，算法治理黑箱的内部构造可被解构为三个层次：界面层和数据层犹如黑箱的皮肤和血肉，而算法所在的模型层则为黑箱注入了灵魂[21]。从过程上看，算法治理通过对日常生活信息的收集、存储、分类与处理，使算法从单纯的计算机代码逐渐转变为一种结构化社会系统的媒介。它由此以黑箱化的方式演化为一种社会基础设施，并不断从算法界面拓展至治理界面乃至社会互动界面，进而推进了算法黑箱向黑箱社会的转型。

（二）敏捷式的界面装置：算法治理黑箱的效能优化

大数据分析模型作为一种治理知识再生产的界面装置，通过将海量数据进行模式化统计分析，使算法获得了预测当前与未来行动的能力。具体而言，大数据分析模型依据标准化原理对数据进行分类、排序与筛选，将一种敏捷式的界面插件嵌入算法治理系统之中。通过将大数据模

型与实际应用管理系统相结合，一系列自动化的数据处理界面装置被嵌入算法治理黑箱之中，赋予其数据高效处理及模型持续优化的功能[22]，实现了算法治理黑箱“存在的编排”。大数据分析模型构建了一套新型的知识再生产体系，从数据采集、组织、存储、处理、共享与利用等环节对数据治理进行系统规范与管理。该模型进一步构建了无缝隙对接的流程界面，在算法治理黑箱内部设置了类似电脑游戏中的层层关卡；任何治理行动必须服从算法预定的标准操作与流程方能顺利通过，否则无法达成预期治理目标[23]。算法治理平台进一步凭借海量的数据资源、日益强大的算力及丰富的算法基础设施，通过预设的算法大模型构建起“输入→执行→输出”的治理闭环，为各类治理问题寻求“最优解”提供了可能。

由此，算法系统逐步承接了国家治理中的决策、执行与监督职责，并在解构传统科层制流程的基础上，建构起一种以速度与自动化为核心的敏捷治理流程。在敏捷范式下，数据标准化技术将原始数据高效转化为“机器可识别”的治理因子，并借助算法软件将其进一步转换为“机器可执行”与“机器可决策”的指令。这一从数据到指令的自动化流水线，极大地革新了治理要素的应用范式，避免了传统治理模式因流程繁琐、响应迟缓而导致的僵化缺陷，从而为算法治理黑箱的界面维护提供了持续且自适应的效能优化机制。

（三）反思性的界面调适：算法治理黑箱的系统庇护

为实现算法治理系统的稳定有效运转，必须在其内部嵌入一种自我调节机制，使其在面对不确定性风险时能够保持相应的韧性。算法治理将数据和模型封装于黑箱的内部界面，同时将一种反思性治理术嵌入系统之中，为算法治理黑箱提供了系统性的庇护机制。

首先，界面的反思性调适为算法治理黑箱提供了技术庇护机制。正如吉登斯所言，反思性监控已成为现代社会系统运转的基础[24]。在算法治理场域，这一特性是指在海量数据的能量支持及敏捷化界面装置的基础上，算法系统及其基础设施能够对自身的治理绩效、决策结果与环境反馈进行持续监测，并据此对其内在的运行参数、价值权重乃至核心决策逻辑进行动态修正。具体而言，算法治理黑箱通过运用海量数据训练机器学习模型，使其在自我调整中实现性能的自我迭代；特别是生成式人工智能所展现的复杂深度学习能力，使得系统能够在最小化人为干预的情况下实现自我学习和进化。正是这种基于反馈的、对自身决策逻辑的批判性审视与调整能力，构成了算法治理的反思性内核，使其能够在动态环境中实现真正的适应性变革，从而为算法治理黑箱提供稳固的系统庇护与持续进化动力。其次，界面的反思性调适为算法治理黑箱提供了制度庇护机制。在算法治理界面的运转过程中，算法体现了人机之间系统控制的规则；算法设计者通过设立标准与目标，使算法系统能够在持续收集实时数据的过程中进行反思性控制，进而在识别、改变及完善界面的过程中实现预先设定的目标[25]。算法治理界面的反思性调适与自动化运转建立在标准与规范的基础上，标准化作为算法治理的制度性基础设施，不仅为界面运行提供了“接口”，也为算法治理黑箱的稳定运行提供了规则保障。

总之，算法治理黑箱的反思性调适，依赖于其技术性庇护与制度性庇护的双重功能。这一过程可被概括为“自适应”与“外干预”相结合的“算法反馈”路径：“自适应”体现了技术庇护下的智能进化，即系统通过数据输入与接收机制来提升数据适应性，实现自我迭代；“外干预”则体现了制度庇护下的规则约束，即通过标准设定、目标校准与干预可见性等机制来规范算法的反馈与决策边界[26]。这种系统的反思性特征，在算法

黑箱的“三阶模式”中得到了充分体现。具言之，算法黑箱在实践中呈现三种类型：基于监督式机器学习技术的初级“黑箱”、依托算法众包模式的中间“黑箱”以及具备自主学习能力的进阶算法“黑箱”[27]。该模式涵盖问题界定、目标／价值选择、数据收集与整合、数据标签、模型构建、应用前评估、应用结果反馈与模型更迭等复杂过程[28]。在此模式下，“自适应”与“外干预”共同作用于数据治理链条的再生产：技术性的自适应确保了治理的敏捷与效能，制度性的外干预则保障了治理的合规与可控。二者共同实现了黑箱的界面反思性治理，为算法治理黑箱的可控性运转提供了动力支撑。

四、界面的固化：算法治理黑箱的两难困境

稳定的治理界面是算法治理系统有效运转的基础，也是算法治理黑箱存在及运行的前提。治理界面为算法治理黑箱进行自我控制提供了媒介，也使其内部结构与外部环境实现适度隔绝。然而，技治主义的界面设计、工具理性的界面规范及内外有别的界面运行方式的共同作用，也加剧了算法治理黑箱的界面固化风险，导致其陷入一种“进退两难”的困境。

（一）技治主义的界面设计：算法治理黑箱的“科林格里奇困境”

技治主义为算法治理的界面设计奠定了控件基础，也为界面运行提供了技术稳固机制。然而，技治主义在塑造治理界面的过程中，也不断嵌入并主导了算法治理的逻辑，使得算法技术逐渐掌控了治理实践的全过程，算法治理的过程性与结果性问题由此被转换为纯粹的技术性问题。具体而言，技治主义的算法设计以明确的治理目标与预设条件为导向，将现实的治理需求转换为代码指令，由此形成一种编程化的界面设计。由于算法治理的输入与输出程序日益脱离人为干预，算法治理界面在技术层面形成了闭环的“黑箱”。这一现象深度契合了“科林格里奇困境”的预言，该观点认为，由于技术发展的后果在前期难以被有效预测，尽管可以对其控制但无法或者没有动力去实施控制；随着技术发展成熟，尽管控制手段增多，但现实中已难以对其后果进行实质性干预[29]。任何一项技术的发展与治理似乎都难以逃出这一困境，尤其随着算法技术将技治主义推向顶峰，这一困境在算法治理场域表现得尤为明显。一方面，算法治理形成了一种人机混合型的决策结构，在政府决策者与算法平台之间建立起多维委托–代理关系，使得算法治理责任呈现折叠性与模糊性并存的状态，从而导致算法治理责任认定面临难以操作的难题。另一方面，算法治理的黑箱化隐藏了其决策过程及机理，不仅导致算法治理面临科学性及“暗箱操作”的质疑，还使得政府对算法的监管手段与领域受到限制。最终，“重机器判断轻人类决策”[30]的技治主义导致了算法官僚主义及黑箱责任监督的双重问题。在此过程中，一种“人造”黑箱在技治主义的界面设计中被不断固化，政府与公众无法清晰掌握算法系统处理日常生活信息的规则依据，也难以对其预设逻辑与事后效果进行有效评估，从而使发展与安全之间的平衡成为算法治理面临的深层挑战。

（二）工具理性的界面规范：算法治理的道德伦理困境

在算法治理界面的建构与运行过程中，数据标准化不仅为其提供了持续运转的“燃料”，也设定了关键的规范机制。标准从其诞生之初就内含着工具理性与价值理性的内在张力。正如有学者指出，标准既能赋予人特定权利，也可为剥夺权利提供合法性依据[31]。因此，理想的数据标准化应力求工具理性与价值理性的平衡，使二者在良性互动中共同

塑造公正的算法治理界面规范。

然而，现实情况却往往背离这一理想设定。数据标准化在强化算法治理工具理性的同时，也遮蔽了其应有的沟通理性。具体而言，工具理性的界面规范体现出“道德物化”的实践逻辑，是工程主义设计理念向日常生活情境拓展的产物。在此机制下，算法治理的黑箱化运作往往将数据标准化单纯视为提升治理效能的工具，而其本应承载的公平正义等价值伦理却遭到忽视。这一工具化导向进一步固化了算法治理的认知边界与操作范围，加深了治理系统中目的与手段之间的断裂。为实现特定治理目标，算法处理器常会有选择性地收集数据，而以何种情境、方式与标准收集数据，必然在事实上影响数据治理的公平性。特别是在多重治理界面中，标准化的工具理性常常会发挥主导作用，进而推动工具主义逻辑与算法治理的黑箱逻辑相互耦合，并在此过程中排斥了标准化治理本应秉持的价值逻辑，由此引发算法治理的伦理隐忧。其结果便是，数据再生产过程中弥散着数据歧视、不公、偏见与排斥等问题，各个阶层在算法治理情境中面临显著的数据鸿沟。最终，工具理性的界面规范将社会既有的偏见与偏好纳入算法系统，并通过数据标准化将其转换为结构性的数据偏见，导致这些失衡的预设立场在算法治理的再生产中被不断延续[32]。在此机制下，算法黑箱已然嵌入并成为社会基础设施的一部分，其运作机制本身构成了一种潜在的“数学杀伤性武器”[33]，系统性地加剧了社会不平等的再生产。

（三）内外有别的界面运行：算法治理黑箱的透明性困境

内外有别的界面运行机制，不仅是算法治理系统稳定的前提，也为算法治理的黑箱化运行提供了空间基础。算法在塑造这种内外有别界面的过程中，加剧了封闭性治理空间的建构，使得“透明性”成为算法治理难以回避的核心困境。算法治理的透明性困境，根源在于其双重依赖性；它不仅取决于算法技术本身的透明性，更与公共治理体系的透明度密切相关。由于算法黑箱与复杂的治理情境存在天然契合性，公共治理部门为了提升治理效能及减少治理成本，具有引入算法治理的强大动力。因此，公共部门与算法平台都倾向于建构内外有别的治理界面，以实现系统的高效运转。正如乌尔曼所言，系统运行的时间越长，参与开发的程序员越多，系统就越有可能变得难以理解，并将在复杂的界面中拥有自己的生命[34]。算法程序的自动化运行进一步强化了这种界面隔离，为社会系统基于规则化理念的运转与算法系统基于专业化的自动化决策之间，埋下了难以兼容的逻辑冲突。

内外有别的界面构造及其运行方式持续提升着“黑箱”的浓度。特别是当算法与科层制结构相耦合时，更呈现出强化界面封闭性的强大动力，进一步固化了算法的排外倾向。在某种程度上，黑箱充当了算法体系运行的一种系统性保护界面；若算法完全透明，则意味着其所有细节，包括用于安全防御的隐藏功能都将暴露[35]。在混沌的治理界面影响下，数据对算法平台及政府部门等治理主体而言日益复杂，对公众而言则越发难以理解。由此，算法治理陷入了治理信息垄断和社会信任脱耦的双重悖论。这种内在矛盾具体表现为：一方面，内外有别的界面构造为信息控制提供了可能，也决定了算法治理无法以真正透明的方式向社会呈现；另一方面，算法治理中的“技术垄断”强调信息集中与流程专业化，这又反过来不断巩固着内外有别的界面结构。最终，这一悖论在现实中引发了系统性的信任危机。从内在的悖论到外化的信任危机，这一演变过程不断加剧着信息控制与社会信任之间的断裂，使算法治理的透明性困境在实践中愈发凸显。

五、界面的治理：算法治理黑箱的规制路径

作为元治理层面的重要范畴，算法治理黑箱的界面治理需要在整合界面要素的基础上，构建算法治理界面的修复、净化与共生机制，从而实现算法治理界面内部与外部的共融共通。

（一）技术道德化：黑箱治理的界面修复机制

对算法治理黑箱的界面规制，不仅需要从技术层面对“黑箱”这一“器物”进行治理，更要求在价值层面对算法治理界面进行系统性重塑。技术折叠是算法治理黑箱生成的重要机理，因此，亟需引入“技术道德化”这一综合性路径来建构界面修复体系，以期打破界面固化的困境。理想状态下的算法治理并非简单地将治理内容通过特定标准转化为冰冷的数据，而是追求工具理性与价值理性的有机统一，旨在实现算法治理界面的价值修复。在此路径下，首先应建立健全道德伦理的嵌入机制。技术道德化意味着对算法治理技术体系进行伦理再造。其核心任务之一，便是将道德要素系统性地嵌入技术架构，从而在算法治理黑箱内部完成伦理价值的有效植入[36]，使算法伦理成为一种约束算法运行的道德评估力量。同时，这一过程不能仅限于技术内部，还必须拓展至社会认知与价值层面。正如研究指出：“人们对于正确事情的信任来源于共享的道德情感基础，但对于一个‘冷冰冰的’、无情的人工智能道德体，却无法对其设置稳定的情感约束。”[37]因此，“道德机器”的目标是使机器在日常生活情境中具备德性判断能力。要走出算法治理中的“科林格里奇困境”，就必须在认知与价值两个维度协同推进现代技术的道德化进程，以协商伦理弥合道德分歧、协同利益冲突，在实现技术与价值的均衡发展中化解算法治理“道德物化”的难题。

其次，应同步建立健全算法治理的“保险丝”机制。该机制是技术道德化理念在风险控制层面的具体实践，要求算法系统不仅要具备效率，更必须具备道德能动性与补救能力。一方面，要求在系统集成过程中构建集场景分析、阈值设定、熔断监测与纠偏修正于一体的界面修复机制，贯彻“事前预防”与“事后维护”有机结合的风险规制理念[38]；另一方面，则需将伦理观念内化于数据治理的技术标准之中，例如在公共数据收集、使用、发布与处理的各阶段，对隐私保护等关键伦理议题进行周密设计，从而赋予治理技术以价值伦理功能。

与单纯利用技术手段消除算法黑箱相比，技术道德化路径为算法治理提供了更深层的价值调控。在工具理性与价值理性均衡发展的视野下，通过综合运用伦理嵌入与“保险丝”机制，方能优化整个治理链条，推动算法治理系统内外价值的融通，最终构建出一种兼具效能与温度的可控治理界面。

（二）适度透明性：黑箱治理的界面净化机制

构建科学、公开的算法治理流程，以提升算法的可理解性与公开性，是算法治理黑箱界面规制的理想目标。在此框架下，透明性包含双重内涵：一是指向内部的可理解性，即通过算法模型的可解释性实现决策逻辑的澄清；二是指向外部的公开性，即算法在采纳、应用与决策等环节的开放程度[39]。然而在实际运行中，算法治理在某种程度上需要通过黑箱的方式才能维持其有效运转；因此，追求算法的完全透明既不现实，也可能会因暴露敏感信息而引发新的风险。基于此，现实的规制方向应从“绝对透明”转向“整体可控”，致力于构建一个“适度透明”的、“可控制”的治理框架[40]。该框架不寻求消除黑箱，而是通过有效的界面净化，在保障系统安全的前提下，促进信息的适度共享与公众理解。

界面净化旨在通过优化界面组件与结构，系

统性提升治理流程的公开性、可理解性与互动性。具体而言，可通过以下三项递进机制实现：首先，建立算法治理信息的差异化披露机制。应以保障政府与公众的知情权及风险防控为基点，根据公共事务的特征，制定分级、分类的算法治理信息公开标准，明确不同场景下的公开内容、范围与方式。在界面设计上，应通过语言规则的简化摒弃官僚化术语并净化互动界面的场景设计，以提升信息的可及性与清晰度。此举旨在解决“公开什么”的问题，以奠定透明的基础。其次，构建算法治理模型的梯度解密机制。为化解技术黑箱带来的公众理解障碍，应构建分层级的解密体系：保障公众对算法基础层的知情权，确保算法治理应用层数据可追溯，同时强化政府对算法核心层的监管，建立起算法治理风险量化评估矩阵。通过构建“算法白箱”，对复杂模型进行降维解构，使公众能够感知并理解其运行机理[41]。此举旨在解决“如何理解”的问题，推动信息从“可见”到“可懂”的深化。最后，完善算法治理的多元回应机制。在披露与解密之外，还需构建一个能对公众质疑进行及时解释与回应的制度体系。这要求制定算法平台公开算法设计、数据来源及处理等方面的标准规范，并通过法律法规明确必须回应的事项，从而在算法研发、决策、应用与监督的全周期内，形成治理主体与社会公众之间的有效沟通与良性互动。

（三）程序正义性：黑箱治理的界面共生机制

对算法治理黑箱的规制是一项复杂的系统工程，需要在算法设计及治理中秉持共生治理的理念，通过制度化的路径，统筹兼顾公共利益与个体权益，最终实现算法的公平正义。公共治理领域的算法应以创造公共价值为最终导向，为实现这一目标，需将共生治理作为总体框架，其内在要求是将公平性约束深度嵌入算法设计与决策之中，并坚持综合平衡的原则，以此重塑治理界面，在共生互动中构建彰显程序正义的算法治理体系。

共生治理意指多元主体在共生界面中以协同行动促进界面各要素的合理流动与有机联结，从而形成一种良性互动的循环发展模式[42]。该理念为算法治理黑箱的规制提供了核心指引和行动模式，其价值集中体现在对程序正义的追求上。此处的程序正义，核心在于保障治理过程的广泛参与性与机会均等性。共生治理体系由共生单元、共生环境与共生模式构成，三者为实现算法治理黑箱的界面共生提供了系统性机制。在共生单元层面，应建立权责明确的多元共治机制。算法正义强调参与平等，即从数据输入到结果输出的全流程都应体现广泛的参与性。因此，需要通过明确算法治理的责任主体、内容及归属，系统界定各方权责、培育责任意识，构建起权责清晰的算法治理制度体系，为多元共治提供稳固的制度基础。在共生环境层面，要打造机会均等的制度环境。算法公平体现为机会平等，是分配正义在算法时代的延伸。这要求一方面通过规范算法处理实现身份中立，另一方面保障社会成员平等享有数字接入权[43]。在此基础上，应将程序正义理念融入算法治理制度设计，建立以数据共享为基础的智能化集成治理系统，并在算法决策、实施与监督全流程中贯彻公平原则。在共生模式层面，需建构算法“向善”的治理结构。“向善”作为一种价值目标，必须通过坚实的规制体系予以保障。这意味着要确保数据标准各环节的公平正义，并构建事前审查、事中监督与事后救济无缝衔接的责任链条。通过内部与外部规制的有机结合，对算法治理界面及全流程实施制度化约束与矫正，从而在自动化决策中趋近算法正义的终极目标。

六、结论与讨论

本文从治理界面视角出发，探讨了算法治理

黑箱的建构类型、生成机理、潜在风险及治理路径，对理解并规制算法治理黑箱提供了新的分析框架。治理界面的混沌性是算法治理黑箱存在的基础，它在驱动算法治理黑箱运转的过程中不断强化黑箱界面边界的封闭性，从而导致界面固化的风险。尽管公众对算法存在诸多疑虑与批判，算法及算法治理黑箱已然深度嵌入社会系统的各个层面，对人类社会产生深远影响。算法治理黑箱作为现代社会技治迷思的产物，集中体现了数字文明变革进程中的风险治理控制难题。因为算法本身无法在复杂社会情境中保持纯粹的技术中立与价值无涉，在现行社会情境下，对算法治理黑箱的完全规制只能是一种理想。现实的治理思路并非要求完全消除算法的黑箱结构，而是倡导通过修复、净化与共生机制的建构，为算法治理黑箱提供一种具有可控性的治理界面。

本文的理论贡献如下：一是初步构建了算法治理黑箱的治理界面框架，从界面的折叠、控制与固化等维度揭示了算法治理黑箱生成的链条机理，为算法治理黑箱的研究提供了分析框架；二是从界面混沌管理的视角阐明了算法治理黑箱的界面层次，辨析了算法治理黑箱的类型结构，深化了其类型学分析；三是从治理界面固化与规制的辩证关系，阐释了算法治理黑箱的潜在风险及治理路径，为相关规制实践提供了理论依据与路径参考。需要指出的是，本文的相关探讨主要是在理论层面展开的，尚未结合深度案例对算法治理黑箱的生成机理进行实证分析。后续研究将在实地调研的基础上，借助典型案例及数据分析，进一步阐释算法治理黑箱的形成路径与规制策略，探索更具操作性与适应性的多元治理路径。

注释：

[1]谭九生,范晓韵.算法“黑箱”的 成因、风险及其治理[J].湖南科技大学学报(社会科学版),2020,23(6):92-99.

[2]BURRELL J.How the Machine“Thinks”:understanding Opacity in Machine Learning Alg-orithms[J].Big data & society,2015,3(1):1-12.

[3]张红春,章知连.从算法黑箱到算法透明：政府算法治理的转轨逻辑与路径[J].贵州大学学报(社会科学版),2022,40(4):65-74.

[4]张爱军,曲家谊.国家治理现代化风险规制:算法中立原则的祛魅与调适[J].理论与改革,2024(4):53-66.

[5]吴椒军,郭婉儿.人工智能时代算法黑箱的法治化治理[J].科技与法律(中英文),2021(1):19-28.

[6]张欣.从算法危机到算法信任：算法治理的多元方案和本土化路径[J].华东政法大学学报,2019,22(6):17-30.

[7]翟翌,罗实.算法行政的合法性危机与化解方案——基于沟通合法性的视角[J].理论与改革,2023(6):63-77.

[8]关晓铭.从算法治理到治理算法：国家治理现代化的技术审视——基于技术政治学的分析视角[J].甘肃行政学院学报,2023(6):49-67+125-126.

[9]张楠,闫涛,张腾.如何实现“黑箱”下的算法治理?——平台推荐算法监管的测量实验与策略探索[J].公共行政评论,2024,17(1):25-44+196.

[10]昌诚,张毅,王启飞.面向公共价值创造的算法治理与算法规制[J].中国行政管理,2022(10):12-20.

[11]SIMON H A.The Sciences of the Artificial[M].Cambridge,MA:MIT Press,1996:6.

[12]李文钊.界面治理分析与发展——构建中国公共管理自主知识体系的本体论框架[J].甘肃行政学院学报,2025(1):1-18.

[13]李春生.技术治理中的算法“黑箱”及其应对策略[N].中国社会科学报,2021-11-10(008).

[14]雷刚.数字政府时代的算法行政：形成逻辑、内涵要义及实践理路[J].电子政务,2023(8):73-89.

[15]Walter W G.The Living Brain[M].New York:W.W.Norton.1953:258.
[16]胡小伟.人工智能时代算法风险的法律规制论纲[J].湖北大学学报(哲学社会科学版),2021,48(2):120-131.
[17]陈鹏.算法的权力:应用与规制[J].浙江社会科学,2019(4):52-58+157.
[18]张海柱.行动者网络理论视域下的算法黑箱与风险治理[J].科学学研究,2023,41(9):1545-1551.
[19]Seaver N.Algorithms as culture:Some tactics for the ethnography of algorithmic systems[J].Big Data & Society,2017,4(2):205395171773810-205395171773810.
[20]伊格纳斯·卡尔波卡斯.算法治理:后人类时代的政治与法律[M].邱遥堃,译.上海:上海人民出版社,2022:36.
[21]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[J].东方法学,2021(4):77-92.
[22]安小米,龙志奇,邝苗苗.标准化视角下大模型数据治理的理论框架及其构成要素研究[J].情报资料工作,2024,45(6):75-83.
[23]王平,梁正.程序员编写代码产生标准?——算法标准在服务过程中进行控制的内在逻辑[J].标准科学,2023(6):6-20.
[24]吉登斯.现代性与自我认同[M].赵旭东,方文,译.上海:三联书店,1998:22.
[25]刘永谋,李尉博.从"大设计"到"小设计":大数据时代的社会规则之变[J].哲学分析,2022,13(1):123-137+199.
[26]宋思茹,洪杰文.数据适应与资本干预:黑箱中的"算法反馈"——基于平台算法工程师的访谈[J].新闻记者,2025(4):18-34.
[27]张淑玲.破解黑箱:智媒时代的算法权力规制与透明实现机制[J].中国出版,2018(7):49-53.
[28]张海柱.算法治理中的不确定风险及其应对[J].科学学研究,2024,42(9):1800-1807.
[29]See David Collingridge. The Social Control of Technology[M].St Martins Press,1980:19.
[30]邬晓燕.数字治理中的技治主义:困境、根源与突破[J].云南社会科学,2024(6):37-46.
[31]Busch L. Standards: Recipes for Reality[M]. Cambridge,MA:MIT Press,2011:17-75.
[32]袁雨晴,陈昌凤.道德物化:大模型人机价值对齐的技术伦理进路[J].南京社会科学,2024(6):88-97.
[33]凯西·奥尼尔.算法霸权:数学杀伤性武器的威胁[M].马青玲,译.北京:中信出版集团,2018:73-79.
[34]Ullman E.Close to the Machine:Technophilia and Its Discontents[M]. San Francisco, CA: City Lights Books,1997:116-117.
[35]张丰羽,汤珂.数字时代的算法滥用及其规制研究[J].经济学动态,2023(2):71-87.
[36]闫宏秀.数据挖掘与技术伦理学的内在路径构建[J].哲学动态,2019(8):95-101.
[37]温德尔·瓦拉赫,科林·艾伦.道德机器——如何让机器人明辨是非[M].王小红,译,北京:北京大学出版社,2017:106.
[38]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[J].探索与争鸣,2025(2):167-176+180-181.
[39]苏宇.优化算法可解释性及透明度义务之诠释与展开[J].法律科学(西北政法大学学报),2022,40(1):133-141.
[40]袁曾.算法应当被解释吗?——人工智能"可控制"的治理向度[J].法学论坛,2025,40(1):130-142.
[41]Arrieta Alejandro Barredo, et al.Explainable Artificial Intelligence(XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI [J].Information Fusion.2020(58):82-115.
[42]何得桂,邓涛.共生治理:中国式城市基层治理创新的新形态[J].黑龙江社会科学,2024(6):24-33.

[43]陈玲,孙君.算法可接受公平:全球算法治理的一个共识机制[J].电子政务,2024(8):2-12.

*基金项目:国家社会科学基金青年项目“基层治理标准化的结构性张力与路径优化研究”(项目编号:22CZZ014)、江西省高校人文社会科学研究思政专项“新媒体场域下社会主流价值理念有效传播路径的研究”(项目编号:SZZX2049)。

作者简介:刘建,江西财经大学马克思主义学院副教授,安徽大学社会治理研究中心兼职研究员,江西南昌,330013;吴理财,安徽大学社会与政治学院教授、博士生导师,安徽合肥,230601。

Black Box and Regulation of Algorithmic Governance: A Perspective Based on The Governance Interface

LIU Jian, WU Licai

Abstract: The governance interface is not only the situational space for algorithmic governance but also a crucial perspective for understanding the black box of algorithmic governance. When algorithms are integrated with the context of complexity governance, they create a chaotic governance interface, shaping a black box structure that integrates technology, cognition, and relationships. High-density data resources provide continuous operational energy for the algorithmic governance black box, while agile interface devices optimize its technical aspects. Reflective interface adjustments enable the black box to self-protect, and these three elements collectively provide a mechanism for controlling the interface of the algorithmic governance black box. However, due to the interface design driven by technocracy, the interface norms guided by instrumental rationality, and the interface operation methods that treat internal and external aspects differently, the interface of the algorithmic governance black box becomes further solidified, leading to a dilemma of being caught between advancing and retreating. Therefore, it is necessary to incorporate interface governance into the discourse of algorithmic governance, regulate the black box through interface repair, purification, and the construction of symbiotic mechanisms, and explore ways to rebuild the trust system of algorithmic governance.

Keywords: Algorithm governance; Algorithm black box; Black box regulation; Governance interface

(责任编校:杨思奇)