计算机科学
COMPUTER SCIENCE

论算法解释权的重构——全算法开发流治理与分级分类解释框架

丛颖男，王兆毓，朱金清

引用本文

丛颖男，王兆毓，朱金清.论算法解释权的重构——全算法开发流治理与分级分类解释框架[J]．计算机科学, 2023, 50(7): 347-354.

CONG Yingnan, WANG Zhaoyu, ZHU Jinqing. Reconstructing the Right to Algorithm Explanation ——Full Algorithm Development Flow Governance and Hierarchical Classification Interpretation Framework [J]. Computer Science, 2023, 50(7): 347-354.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900120

论算法解释权的重构

——全算法开发流治理与分级分类解释框架

丛颖男[1]　王兆毓[2]　朱金清[3]
1 中国政法大学商学院　北京 100088
2 清华大学法学院　北京 100084
3 北京字节跳动网络技术有限公司　北京 100043
(cyn_2010@163.com)

摘　要　随着人工智能技术的快速发展,自动化决策算法逐渐进入公共领域并越来越多地影响到社会公益与个人权益。而相应的算法风险如算法歧视、算法偏见、算法垄断等不断出现,进而引发了算法治理的切实需求。面对自动化决策使用者与用户之间信息、技术的不对称地位,传统法律资源不敷适用,对自动化决策用户保护之权利的不足成为算法解释权的必要性基础。作为算法治理的重要手段,算法解释权的价值在于对算法的"黑盒"构建"适度透明性",矫正开发者与用户之间的信息不对称,并且再平衡双方畸形的分配风险负担,成为规制自动化决策使用者、保障用户权益必不可少的制度配置,因此算法解释权研究成为国内外学界与司法实践共同关注的焦点。而在现行法视野下,算法解释权制度存在适格主体过于狭窄,保护范围不够全面,权利内容尚需明确等问题。对此,在解构算法解释权的基础上,从全算法开发流治理与分级分类解释框架的视角对算法解释权制度进行重构。通过全算法开发流治理的建构,对算法解释权的主客体进行适度扩张;通过分级分类解释框架的构筑,结合个案视角明确算法解释权的内容与边界,以此兼顾算法的个性与共性,平衡算法解释的效率与用户权益的保护,全面保障自动化决策中的各方权利主体利益,为数字经济发展赋能。

关键词:算法解释权;算法治理;自动化决策;个人信息保护
中图法分类号　TP182

Reconstructing the Right to Algorithm Explanation

—Full Algorithm Development Flow Governance and Hierarchical Classification Interpretation Framework

CONG Yingnan[1],WANG Zhaoyu[2] and ZHU Jinqing[3]
1 Business School,China University of Political Science and Law,Beijing 100088,China
2 School of Law,Tsinghua University,Beijing 100084,China
3 Beijing Bytedance Network Technology Co.,Ltd,Beijing 100043,China

Abstract　With the rapid development of artificial intelligence,automated decision-making algorithms(ADM) have gradually entered the public domain and increasingly affected social welfare and individual interests. Meanwhile,emerging risks of ADM,such as algorithmic discrimination,algorithmic bias,and algorithm monopoly have raised the demand of governance to algorithm. Faced with information and technology asymmetry among parties involved,traditional legal resources fall short in protecting the rights of users in ADM,which justifies the right to explanation. In addition,the right to algorithm explanation,serving as an important means of algorithm governance,is conductive to making the black box of algorithm moderately transparent,correcting information asymmetry,and balancing the risk burden between the deployer and the user. It has thus become a necessity in regulating ADM deployers and safeguarding the interests of its users. Therefore,the right to explanation has become the focus in both academic and practical realms from home and abroad. However,the right to algorithm explanation in China is faced with the problem of limited eligible parties,insufficient protection scope,and inexplicit content of rights. In this regard,this paper advocates deconstructing the right to explanation and further reconstructing it from the perspective of machine learning workflow with a hierarchical classification framework. Introducing the concept of machine learning workflow can reasonably extend the scope of the subject and object of the right,while establishing the framework of hierarchical classification can clarify the content and boundary of the

基金项目:北京市教改项目"法商大数据分析创新型人才培养模式研究"(京教函[2020]427号);中国政法大学新兴学科培育建设计划
This work was supported by the Beijing Education Reform Project "Research on the Training Mode of Innovative Talents for French Business Big Data Analysis"(Jingjiaohan [2020] No. 427) and Cultivation and Construction Plan of Emerging Disciplines of China University of Political Science and Law.
通信作者:朱金清(zhujinqing@bytedance.com)

right, which considers both individuality and generality of algorithms and balances the efficiency of explanation and the protection of users' rights. In this way, all parties in ADM can be fully protected, and the development of digital economy can be empowered.

Keywords Right to explanation, Algorithm governance, Automated decision making, Protection of personal data

1 引言

算法是"一个人工定义的、输入并输出单个或多个值的计算过程"[1]。只要满足输入、输出、明确性、有限性和有效性要素的运算过程,都可以被称为算法[2]。随着人工智能技术和大数据科学的发展,算法已经在定位导航、外卖配送、自动驾驶、个性推荐、智能投顾等许多场景下得到广泛的应用,自动化决策带来了效率优化、流程简化等实益。更有甚者,算法已经不断进入公共领域并与公民的权利义务息息相关。算法在犯罪风险预测、离婚案件财产分割、交通违章处理、公共福利计算[3-6]等许多司法、行政场景下对公民权利之处分起到了辅助甚至决定决策的作用。故而在诸多应用领域之中,算法俨然已成为一种"准公权力"[7]。如果算法已经从纯粹的商业私人领域走入涉及国家安全、公民基本权益等公共利益的公有领域,此时的算法绝非单纯的技术工具,而是潜移默化地嵌入公权力服务并成为其重要部分[8]。这将最终导致一个规模空前的"评分社会"抑或等级化森严的"排序社会"正在逐渐成型[9]。

然而,在算法的自动化决策高速发展的同时,其外部性风险亦逐渐浮出水面。"越来越多的人认为,算法使不平等现象自动化,算法是复制种族主义的有偏见的黑盒子,或者算法控制了我们的金钱和信息"[10]。算法偏见、算法歧视、算法操控、恶意用户画像、信息茧房与信息回音室等算法风险不断出现。而应用于公共领域的算法,其缺漏则会对公民的基本权利产生直接的侵害。在刑事司法中,美国部分州使用 COMPAS 系统对公民的犯罪风险进行评估并将其应用于定罪量刑之中[11]。公民 Loomis 被系统评估为"高风险",并最终因偷窃和拘捕被判刑六年,而后其提起上诉,认为威斯康星州法院对 COMPAS 系统的使用违反了正当程序原则。在社会保障中,爱荷华州利用算法计算居民的救济金,许多公民认为该算法对其应得的救济金计算错误,严重损害了其根本利益并对其提起上诉[12]。研究表明,人工智能系统在进入大规模应用之前,只有 30%的开发者对其算法进行了详尽的检验[13]。并且,随着算法应用深度与广度的不断增加,其外部性风险亦与日俱增。推荐算法为信息捕猎者提供了便利,影响了政府决策,部分算法风险识别工作本身就交给了算法,甚至算法利用人的生物弱点与内心欲望来控制"自由意志"[10]。传统决策的决策约束程序如通知、公告、评论、参与、回避、异议、救济等,旨在避免决策武断和恣意,保障决策可信和正当的制度架构,在算法决策面前均一定程度失效。治理层面尚未探索出一套行之有效的机制,来确保建立治理层面的技术信任。如若不对算法进行合理有效的治理与规制,"在算法侵入前,时间不多了"[14]。在此背景下,算法规制与算法治理已经成为我国学界和司法实践众所瞩目的研究方向。学者们已经对算法反垄断、算法治理、算法歧视等多个论题展开了广泛的讨论并给出了可行的解决路径[15-17]。与此同时,在立法中,2021 年 8 月《个人信息保护法》(以下简称《个保法》)的出台为自动化决策的法律规制奠定了体系性的基础。而 2021 年 9 月《关于加强互联网信息服务算法综合治理的指导意见》(以下简称《治理意见》)和 2021 年 11 月《互联网信息服务算法推荐规定》(以下简称《推荐规定》)相继出台,这两部直接以"算法"命名的法律文件选取了信息服务这一典型自动化决策场景,提出了算法规制的"中国回应"。

在算法治理与算法规制体系之中,算法解释权"已经成为学界共识并在多国实践中展开",并且被视为"算法决策规制的中心"[18]。源自域外的算法解释权制度成为比较法与现行法、法学与信息科学的连接点,并成为算法治理研究的前沿问题。然而现行法视野下的算法解释权制度仍存在着诸多不足,本文在解构算法解释权的基础上,从全算法开发流治理与分级分类解释框架的视角切入,以算法自动化决策涉及的各方权利主体利益为中心,对算法解释权的重构与完善进行了探索,提出了算法治理必需的司法保障与相关技术,为数字经济稳健发展赋能。

2 算法解释权的权利解构

2.1 算法解释权的价值基础

在讨论算法解释权的权利结构之前,首先需要解决的问题是,算法为何需要解释?

算法解释权的价值在于其无法替代的制度优势。其一,算法解释的作用为在算法的"黑箱"中构建适当的透明度,从而成为自动化决策使用者和用户之间信息不对称的有效纠偏工具。其二,自动化决策使用者与用户之间基于意思自治达成用户协议,使用者对信息、数据的支配地位导致算法解释权是协议必然衍生的权利,也显然为协议的应有之义。其三,算法解释权也是对合同风险的合理分配,自动化决策事实上的强制性和算法开发者、使用者具有的强大风险控制能力,导致处于相对弱势的用户应当获得倾斜性保护,通过算法解释权可以平衡此畸形的风险负担[19]。

在算法解释权本身的价值之外,创设算法解释权制度不得不回应的问题是,自动化决策双方通过意思自治允许自动化决策算法之运行,为何要超越合同配置给自动化决策相对人以额外的权利?这是因为在自动化决策合同双方民事主体在法律意义上平等的伪饰背后,是自动化决策使用者和其相对人不平等的事实地位。自动化决策使用者处于信息、技术、市场的支配地位,不论在风险预估还是风险规避上都有其绝对优势,而现有法律制度则难以作出回应。首先,用户注册平台账户时必然需要对"用户协议"进行授权,这一强制同意使得自动化决策使用者进行的信息收集、数据处理的任务合理化、合法化,甚至可能借此规避其非法收集、处理行为的法律责任。其次,现有法律责任的认定规则难以对自动化决策

造成的可能损害进行界定。例如，“用户画像”等算法应用是否违法，在司法实践中难以判断。而即使该行为被认定为违法，算法使用者也可就保护商业秘密为由抗辩。这最终会导致当事人因自动化决策的侵害而寻求救济时，不论是从民事合同效力瑕疵的角度主张重大误解、显失公平、欺诈等可撤销条件，从侵权责任体系下主张损害赔偿请求权，抑或主张消费者知情权，均无法充分保护其权利并得到有效的救济[10]。因而在算法解释权的视角下，传统的合同制度需要做出因应性的调整[19]。创设算法解释权的根本目的即在于构建自动化决策双方的实质平等，为自动化决策的弱势一方——用户，提供倾斜性保护与额外救济。在大数据、人工智能技术高速发展的数字经济时代背景下，算法解释权已然成为算法公平、决策正义的必然要求。

2.2 算法解释权的起源

回溯算法解释权之起源，该制度最早起源于欧洲。早在1972年，法国颁布的《关于信息技术、数据文件和公民自由的信息自由法案》即对算法解释进行探索[20]。该法案规定，数据主体在一定前提下有权知晓对其数据进行处理的算法逻辑，并可以拒绝算法处理。1978年颁布的《法国数据保护法》在此基础上规定，数据主体在不违反版权规制的前提下，有权获得算法运行相关的规则逻辑并提出质疑[21]。1995年，欧盟颁布了《数据保护指令》(以下简称《指令》)(Data Protection Directive，DPD)。《指令》第十二条提出，数据主体有权了解对其个人数据的处理方式以及处理规则。而随后，欧盟于2016年4月颁布的《通用数据条例》(General Data Protection Regulation，GDPR)建立了对数据及数据利用的全方位规制体系，并成为世界数据立法参考的对象[10]。GDPR第十三条沿袭了《指令》的相关规定，即数据主体有权了解自动化决策及与其相关的设计逻辑、风险后果等(见GDPR Art. 13.2(e)与Art. 13.2(f))。第二十三条也提出了数据拥有者需采取相关措施保证数据主体对自动化决策的知情权和质疑权见(GDPR Art. 23)。同时，GDPR的序言(Recitals)部分中，第七十一条明确提出了解释权，并规定了决策相对人请求对决策解释的权利(见GDPR Recital 71.)。但有学者指出，上述条文不具备事实上的法律强制性，且算法解释的具体内容只是算法的一般功能、实现目标等，最终导致算法解释权形同虚设[22]。

而美国早在1986年的《电子记录系统和个人隐私》评估报告中即对算法规制进行了尝试。该评估报告明确提出，自动画像技术可以通过算法归纳等技术拟合个人的行为模式，政府部门对此类技术的不当使用可能会引发对公民隐私权的侵犯，但当时对个人画像的法律规制尚为空白。在后续评估报告与议会听证的基础上，联邦参议员Zablocki提出了《公平信贷报告法》议案。该议案赋予消费者在受到负面征信报告影响时，知晓作出该负面报告的事实依据的权利。这是较早在征信报告的场景下，公民获得算法解释权的立法尝试之一[23]。而后，美国联邦贸易委员会(Federal Trade Commission，FTC)分别于2020年和2021年发布了《人工智能和算法运用》(Using Artificial Intelligence and Algorithms)及《商业人工智能应用的真实性、公平性与平等性目标》(Aiming for Truth，Fairness，and Equity in Your Company's Use of AI)，对算法解释的原则、标准、内容进行了指导。目前，算法解释权的重要性已经在立法者与学者间获得了广泛认同，自动化决策使用者有义务对算法进行合理的解释，自动化决策用户有权利知晓算法的部分训练过程、运行逻辑与潜在风险。

2.3 现行法视野下的算法解释权

《民法典》《个保法》与《推荐规定》共同构成了我国现行法体系下对于算法解释权体系的规制格局。首先提出算法解释权的是《民法典》，《个保法》则进一步确立了个人信息处理者对信息处理目的、方式和范围的告知说明义务。《民法典》第一千零三十五条规定了个人信息处理者“明示处理信息的目的、方式和范围”的义务。《个人信息保护法》第四十八条规定：“个人有权要求个人信息处理者对其个人信息处理规则进行解释说明。”虽然此处的“信息处理规则”不能直接等同于“算法”，但算法事实上主导了数据收集、处理和输出的全生命周期，因而在大部分场景下，对“信息处理规则”的解释即为对“算法”的解释；其二是《个保法》强化的自动化决策的透明度、公平性原则和“重大影响”下的解释说明义务，《个人信息保护法》第二十四条规定：“个人信息处理者利用个人信息进行自动化决策，应当保证决策的透明度和结果公平、公正，不得对个人在交易价格等交易条件上实行不合理的差别待遇……通过自动化决策方式作出对个人权益有重大影响的决定，个人有权要求个人信息处理者予以说明，并有权拒绝个人信息处理者仅通过自动化决策的方式作出决定。”最后，在《个保法》基础上，《推荐规定》针对信息服务推荐算法这一特定使用场景，对算法推荐服务提供者规定了算法解释义务。在此基础上，有学者对算法解释权做出了如下定义：当自动化决策对其用户造成显著影响时，用户有权向自动化决策使用者提出异议，要求其提供解释，并且有权要求更新数据或更正错误[24]。在此定义基础上，笔者认为，现行法视野下的算法解释权的主要特征如下。

其一，就权利的性质而言，算法解释权显然属于请求权之范畴，且请求权的主体仅限于用户[18]。而用户在自动化决策中通常处于信息、技术、知识的弱势地位。这就导致即使用户提起了算法解释请求权，自动化决策使用者对其算法进行解释后，用户也可能不具备理解该算法解释的知识储备，不具备检验算法解释真实性的客观条件。若仅由用户作为算法解释请求权的适格主体，即使开发者从技术层面进行了充分的算法解释，于用户而言算法之透明度并不能当然地增加，且该算法解释的充分性与真实性也难以得到检验，无法从实质上保障用户的权益。

其二，就权利的行使条件而言，当且仅当自然人认为自动化决策“对用户权益有重大影响”时方可提起算法解释请求权。有学者认为，对于算法的时机而言，算法解释不仅可以是对决策内容的事后解释，也可以是对系统功能的事前解释[25]。但在现实场景之下，直至算法作出决策之前，用户常常并不知晓该算法的存在，更不可能基于“对用户权益有重大影响”提起算法解释请求权。因而现行法视野之下的算法解释权，是一种事后的权利，不能在事前、事中保障用户的合理权益。法律不应只是事后追责，更是事前防范与事中监管，

仅仅通过事后权利进行救济，将会使自动化决策的用户陷于潜在的风险与侵害之中。

其三，就权利的内容与标准而言，现行法视野下对算法解释提出了“透明度和结果公平、公正”之标准。但结果公平、公正如何判断？算法透明度达到何种程度？算法解释之内容又具体为何？这是司法实践所面临的且亟待解决的问题。

如上所述，笔者认为现行法视野下的算法解释权存在着适格主体过于狭窄、权利时效过于滞后、解释内容与解释标准过于模糊等问题。为了充分保护自动化决策用户的合法权益，同时兼顾自动化决策使用者的利益，有必要对其进行适度重构。在算法解释权的本体论角度上，应当树立全算法开发流治理的算法规制视角，在此基础上对算法解释权的客体进行适度重构，而客体的重构也对算法解释权的权利主体和权利义务的重构提出了必然的要求；在算法解释权的方法论角度，可以从分级分类解释框架的构建，针对具体算法采用个案视角的方式进行解释。

3 全算法开发流治理的重构

重构算法解释权，起点是对算法解释权主客体的再次思考。而这一思考的前置条件，则是对算法本质的再认定。目前，在数据法学领域，数据“全生命周期”保护的理念已经得到法律和政策的落实。《数据安全法》第三条从数据生命周期的角度将数据处理行为定义为：“数据处理，包括数据的收集、存储、使用、加工、传输、提供、公开等”。此外，2021 年 7 月，工信部发布《网络安全产业高质量发展三年行动计划（2021—2023 年）（征求意见稿）》。征求意见稿指出，“要加强数据全生命周期安全保护，保障人民群众的生命财产安全和个人隐私安全。”数据不仅仅是静态的二进制字符，数据在其全生命周期中呈现出不同的特征与状态，这对数据保护的全面性和不同阶段之间的差异性提出了要求。对算法的认定也应考虑上述情况，不应局限于算法这一开发流程的结果，而应从“全算法开发流”的治理视角出发，对算法解释权进行本体论角度的重构。

3.1 全算法开发流的视角的构建与算法解释权的客体

目前，对于算法的通用开发流程，已有了标准化的总结与探索。Saleema 研究团队[26]对数据挖掘和数据科学中常用的工作流如 Team Data Science Process（TDSP）[27]，Knowledge Discovery in Databases（KDD）[28]和 CRoss Industry Standard Process for Data Mining（CRISP-DM）[29]等进行分析，针对机器学习的发展现状，对传统的算法开发流进行了新的规划与定义，亦为人工智能时代下算法解释权制度的构建提供了借鉴意义。因而本文在 Saleema 的研究成果基础上，尝试对“全算法开发流视角”进行讨论。

Saleema 将通用的算法的全开发流程分为以下 9 个具体阶段：模型需求（Model Requirements），数据收集（Data Collection），数据清洗（Data Cleaning），数据标签（Data Labeling），特征工程（Feature Engineering），模型训练（Model Training），模型评估（Model Evaluation），模型部署（Model Deployment），模型监督（Model Monitoring）。在模型需求阶段，开发者需要定义机器学习可以实现的功能，并思考这些功能在具体产品中的应用方式。在本阶段中，开发者还针对给定的任务，确定具体的模型类型。在数据收集阶段，开发者找寻现有的数据集或进行数据集的开发。通常，开发者会根据通用数据集进行预训练，而后采用迁移学习[30]等技术，并结合专有领域的数据进行改进。数据清洗阶段是各种数据活动中的常用步骤，主要用于清洗数据中的错误数据或噪声数据。数据标签阶段需要对数据中的相应内容打上标签，这也是监督学习中对数据集的基本要求。微软各团队的数据标注工作通常由工程师、行业领域专家或众包平台完成。特征工程阶段旨在为模型训练选取合适的特征。但对于 CNN 等模型来说，这一阶段通常与下一阶段的模型训练同时进行，特征工程这一步骤时常不具备独立性。而后在模型评估阶段，开发者通过测试集对模型的输出效果进行验证并检测模型的泛化能力。在某些领域中，这一阶段可能还需要大量的人工校验。模型通过评估之后，将在模型部署阶段被部署到具体的产品或设备中，并在模型监督阶段中对其效能进行持续的监督与反馈。

笔者认为，算法解释的客体应当为自动化决策算法的“全算法开发流”。即在进行算法解释时，所关注的客体不应当仅为算法开发流的产物——狭义上的“算法”，而应当构建全算法开发流的解释视角，算法解释应当兼顾算法全开发流的各个阶段以保障其全面性，同时又应当对各阶段进行更细致的规制以确保其差异性。具体而言，目前的算法解释常常只限于模型训练阶段。然而，在自动化决策开发流前期，模型需求、数据处理阶段以及后期模型验证以及模型监管阶段对用户可能产生的侵害不可忽略。“全算法决策开发流”的治理视角要求算法解释的客体不仅针对数据标签、特征工程、模型训练等传统意义上的算法开发阶段，还包括事前对模型需求、数据收集、数据清洗阶段的解释和事后对模型应用以及模型监测的解释。由此方可对自动化决策开发的全流程进行有效的法律规制，充分保障用户的权利运行。

3.2 全算法开发流下的权利主体

全算法开发流治理的重构首先要求了对算法解释权客体的重构。在此基础上，算法解释权的主体也应当进行因应性的调整。现行法视野下，算法解释权的主体为“对个人权益有重大影响”的自动化决策用户，如在贷款中被预测为不具备还款能力而申请贷款失败的银行客户、在就业时被预测为不具备工作能力而被拒绝的求职者等。然而，在“全算法开发流”解释的逻辑之下进行思考不难发现，如若自动化决策的相对人是算法解释权唯一的适格主体，则“全算法开发流”的解释将失去意义。这是因为在算法投入实际应用之前，自动化决策的相对人对其并不知晓，更不可能借此认为其受到自动化决策的不利影响。例如，在数据收集阶段，存在着算法过度收集个人信息而对用户权益产生侵害的情形。在此种场景之下，相对人对自动化决策可能的不利影响难以作出准确预期，提起相应的算法解释权亦如空谈。

因而，“全算法开发流”的解释必然要求算法解释权适格主体的有序扩张。笔者认为，算法的适格主体，应当在一定条件下吸纳公权力机关与第三方监管机构，如行业协会、第三方评估机构等。

适格主体扩张之必要性，一是在于上述“全算法流程”解释的必然要求，二是在于公权力机关监管与第三方监管各有监管优势，有其提起条件。倘若赋予公权力机关与第三方机构过于宽泛的请求权提起条件，可能会导致算法解释请求权的滥诉，进而对算法技术开发与发展产生负面影响。笔者认为，只有针对直接应用于公共领域的算法，在公权力机关的职权范围内，公权力机关有权利也有相应的义务对其进行监管，在此场景下公权力机关的适格主体资格来源于保障其公权力运行自然而然衍生的监管要求。而对于第三方监管机构而言，算法使用者只有在具有较高的解释义务之时，即下文所述“高风险”算法的使用者，具有提请算法解释权之主体资格。第三方监管机构的监管优势来自于其技术之专业性，其主体的适格性来源于用户和公权力机关在技术方面的不足。在“高风险”算法之中，对个人用户抑或公权力机关的解释并不能真正从技术上排除算法风险，因而有必要通过专业的技术监管请求算法使用者对其算法进行详尽解释。同时在此场景下，第三方监管机构也具备一定的保密义务。例如，有学者主张通过“公开算法模型”的方式进行算法解释[8]，也有学者从商业秘密保护、算法安全和用户理解能力的角度认为对算法开源并不可取[10]。而此时，通过引入第三方监管机构，可以通过仅对第三方机构开源并赋予其审查义务与保密义务的方式解决上述分歧，“全算法开发流”视角下算法解释权的权利主体重构有显著的制度优势。

3.3 全算法开发流下的义务主体

在通过“全算法开发流”的治理思路对算法解释权的客体与权利主体进行重构之后，如何重构全算法开发流下的义务主体亦值得思考，算法解释权的义务主体亦需要进行因应性重构。

重构算法解释权之义务主体，首先需要考虑的背景是，自动化决策算法的开发者与其使用者常常不一。随着算法规模、复杂度的不断增加，部分算法使用者不具备相应的技术能力或开发时间，采用“外包”“转包”等形式，委托第三方对其算法进行开发。而后对于第三方开发的算法，对其进行产品验收后直接部署至业务场景，或进行简单的再开发或迭代后投入使用。开发者拥有技术优势，而使用者则常常拥有商业地位优势，二者相互结合，共同完成算法的开发与部署。在开发中，使用者提出其需求，开发者进行实现；在部署中，开发者将其算法部署于使用者的应用环境之中。

在现行法视野之下，自动化决策的义务主体为“对个人权益有重大影响的”自动化决策的使用者。实际上，在现实中，由于自动化决策使用者常常为网络服务平台，将其作为自动化决策义务主体在实质上反映了网络服务平台的“平台责任”，例如使用征信算法评估客户还款能力的银行、通过定损算法进行赔付的保险公司、通过个性推荐算法吸引用户的视频平台等。平台直接利用算法完成其使用场景下的特定功能，直接影响着用户的权利，其义务主体资格似乎毋庸置疑。这一规定有违“技术中立”之理念，但事实上，现代互联网环境下的“技术中立”已经饱受质疑[31]。当算法频频对用户产生不利影响时，似乎技术中立之理念已难以被法律恪守，除了算法的开发者之外，平台也应当对其在算法部署与算法应用等方面的过失承担责任[24]。对于自动化决策使用的平台责任，其有着合法性与正当性基础。

而在上述算法开发者与使用者二分的场景下，算法使用者由于技术水平等因素，未必能真实、有效、全面地向自动化决策相对人进行算法解释。因而笔者认为，在此种情况下，算法的实际开发者虽然不是算法解释义务的直接主体，但有协助算法使用者进行算法解释的义务。因而，算法开发者成为事实上算法解释权可能的义务主体，辅助算法使用者实现其算法解释义务。

4 分级分类解释框架的重构

本文通过引入“全算法开发流”的治理视角，在本体论层面对算法解释权的主客体进行了重构。而在主客体之余，算法解释权的内容与算法解释的方法尚未得到现行法律法规的明确回应。为了从方法论角度对算法解释权进行重构，系统性地对算法解释权的内容与解释方法进行了合理有效的规制。笔者主张，为了在提高规制效率的前提下尊重不同算法的个性差异并建立算法解释权的差序规制格局，应当在尊重个案解释的基础上，建立算法分级分类解释框架，并对算法解释的边界进行明晰。

4.1 算法分类分级解释框架的构筑

分类分级的规制路径已经在我国现行法中广泛体现。《数据安全法》和《个保法》在对数据、个人信息进行规制时，均采用了分级分类规制的手段。《数据安全法》中第二十一条明确提出了“国家建立数据分类分级保护制度，根据数据在经济社会发展中的重要程度，以及一旦遭到篡改、破坏、泄露或者非法获取、非法利用，对国家安全、公共利益或者个人、组织合法权益造成的危害程度，对数据实行分类分级保护。”同时界定了核心数据、重要数据的分类基准。《个人信息保护法》第五十一条也明确提出了“对个人信息实行分类管理”的规定。在算法规制中，分级分类规制的理念也得到了确认，《推荐规定》第二十三条规定：“网信部门会同电信、公安、市场监管等有关部门建立算法分级分类安全管理制度……对算法推荐服务提供者实施分级分类管理”。笔者认为，算法的分类分级制度代表了对算法从横向与纵向结合的双重分类视角。分类代表了算法在不同应用领域内的划分，分级则解释了在同一应用领域下不同算法的差异，二者相辅相成。

在算法的分类解释中，由于算法的应用领域常常是其所依托平台的外化，因而算法分类亦可以参考平台分类的相关规定。2021 年 11 月由国家市场监督管理总局发布的《互联网平台分类分级指南（征求意见稿）》（以下简称《分类分级指南》）即对平台分类进行了尝试。《分类分级指南》提出了网络销售类、生活服务类、社交娱乐类、信息资讯类、金融服务类、计算应用类的分类标准。而在每一大类中又细分了具体子类，例如在网络销售类平台中，又具体分为综合商品交易类、垂直商品交易类、商超团购类；在生活服务类平台中又具体分为出行服务类、旅游服务类、配送服务类、家政服务类、房屋经纪类。对于不同类别的算法，应当动态确定其解释内容的框架，并根据算法所处的层级进行规划。例如，综合商品交易类算法属于网络销售类算法的子类别，因而在对其解释时，应当

根据网络销售类算法的解释框架，结合综合商品交易类与垂直商品交易类、商超团购类算法的差异进行配置。

而在算法的分级解释之中，GDPR 较早针对算法风险提出了分级制度，分为不可接受的风险、高风险、优先风险和低风险 4 类。在我国，深圳市于 2021 年 7 月发布了《深圳经济特区人工智能产业促进条例(草案)》(以下简称《促进条例》)。《促进条例》明确提出了对人工智能算法依据其风险等级实施分级分类差异化监管，并将人工智能算法分为高风险与中低风险算法，采用不同的监管监控模式。在此基础上进行思考，笔者认为在对算法进行风险评估与算法解释时，可以依据其风险将其分为极高风险算法、高风险算法、中风险算法和低风险算法 4 类，并对 4 类算法的算法解释提出差异化的规制要求。

4.1.1 极高风险算法

极高风险类算法直接应用于司法、行政等与当事人的基本权利息息相关的领域，对当事人的核心权利有直接处分作用。由于现在主流的大数据驱动的深度学习、神经网络模型的本质是建立输入输出之间的相关性，该类模型并不具备真正意义上的可解释性[32]。也就是说，该类算法的风险永远无法真正消弭。另外，在此种场景下若采用自动化决策算法，显然需要对其提出极高的监管、规制需求，最终可能导致进行自动化决策的效率与成本反而不如人工决策。因而笔者认为，在此类场景中，算法的应用范围应当有所限制。类似于“法律保留”要求了限制人身自由的处罚等只能由法律设定，在风险极高的应用场景下也应当“算法保留”——算法只能起到辅助决策的作用，不能替代人类直接进行决策。这一观念也被广泛接受，美国威斯康星州要求法院适用算法量刑时需要保证人类参与到实质决策中[33]，我国学者也认为在智能裁判等领域中算法永远不能替代人类法官而只具有辅助决策之地位[34]。

4.1.2 高风险算法

高风险算法“对用户权益有重大影响”，并且该可能造成的风险不能显著大于自动化决策的效率收益。笔者认为，在此类场景下，可以适用自动化决策，但需要对其进行非常审慎的规制。

具体来说，对于高风险算法，其算法解释权的行使主体可以包括上述第三方机构、用户、公权力机关各个主体；并且算法解释不仅是事后解释，也包括事前对算法可能风险的评估、算法实现路径的备案等算法解释手段，还包括事中对算法训练、算法验证的动态报告等解释措施，以此实现对高风险算法的全方位监管。

例如，对个人信息进行大规模处理的算法显然应当落入高风险算法之范畴。个人信息权全是保障自然人人格尊严的重要权利，倘若不对个人信息利用的算法加以谨慎的规制，不免产生信息利用的“丛林法则”和“公地悲剧”[3]。算法权利与个人信息权常常深度融合[36]，严格的算法解释规制是对算法权利和个人信息权的双向保障。

4.1.3 中风险算法

中风险算法能在一定程度上影响当事人的权益，并且在某些场景之下可能“对用户权益有重大影响”，但其潜在的风险远小于自动化决策带来的收益。笔者认为，对于中风险算法，算法解释权的配置按照我国现行法的规定配置足矣，即仅可由用户提起的事后解释。

例如，新闻推荐场景下，新闻推荐者通过信息控制另一方主体获取信息的渠道和程度，同样有可能对用户权益产生影响[37]。但总体而言，新闻推荐算法给用户造成侵害的可能性与损害程度均在可以控制的范围内，并且可以被新闻推荐算法自动化带来的便捷所稀释。

4.1.4 低风险算法

低风险算法并不具备直接处分当事人权益的可能，因而跳脱出“对用户权益有重大影响”之范围，也自然不需要利用算法解释权对其进行规制。通过低风险算法这一级别的配置，可以有效限缩用户提起算法解释权的范围，防止“滥诉”，从而兼顾算法开发者的基本利益，促进算法技术的稳健发展。

4.2 分类分级解释的个案视角

算法分级分类框架的目标在于促进算法解释权制度之效率，对同一级别、同一类别的算法，在解释算法时采用相对统一的框架进行标准化、高效化。然而法律规制的目标，在共性之外，显然也需要兼顾个案的个性，进而避免算法开发者超出合理范围的解释义务，保障算法解释权的实现具有灵活性。本文认为分级分类解释框架的制度定位是在算法解释中实现总体的解释效率，而个人视角则让算法解释在个案中实现解释正义。

在个案视角下，应当为算法解释确定基本的解释范围，并动态配置算法解释者的解释义务。根据解释内容的不同，可以将算法解释分为“内部解释”和“外部解释”[38]。“内部解释”面向的对象通常是具备专业技术素养的开发人员，解释算法设计的具体细节内容包括但不限于数据集选取、数据清洗规则、模型指标选取、模型设计、训练方式、验证方法等。通过此类专业性的解释，可以从技术角度阐述算法对所需功能的实现程度，并对可能的风险产生预期。而“外部解释”面向的对象通常是不具备专业算法知识的公众等。外部解释要求通过公众能够理解的方式对内部解释的内容通俗化、可理解化，即“对算法解释的解释”，用以阐释算法实现的目标、大致的实现路径以及相应可能的侵害，并说明算法开发的合法性与合规性。据此，有学者对其进行总结，认为内部解释的解释标准应为“可判断性”(Interpretable)[39]，而外部解释的标准则为“可理解性”(Comprehensible)[40]。

结合上述算法分级分类解释带来的解释权主体的差异，可以确定算法个案解释的基本方法。即对于高风险算法来说，当面向第三方机构进行解释时，第三方机构有相应的技术能力理解算法的底层运行基础，因而应当对其采用“内部解释”，在风险较高的场景下甚至可以通过向第三方部分开源保障其监管的全面性。而不论是高风险算法抑或低风险算法，当面向用户与公权力机关进行解释时，考虑到其技术方面的认知能力，应当进行“外部解释”。

4.3 算法解释的边界

在算法实践中，算法解释存在很多障碍与困难。而纵使算法解释权制度的初衷是为了保护处于不利地位的用户，其也绝不意味着绝对的透明与公开，并不能借此扩张算法解释权的

边界，因而法律也应当对算法解释的边界作出回应。笔者认为，算法解释权的边界来源于技术的局限，例如模型可解释性的缺失，也来源于国家安全、商业秘密的限制。

首先，模型可解释性的限制导致了算法开发者并不能从真正意义上解释算法运行的内核，对算法技术进行完整解释亦不能从根本上杜绝算法风险。但在算法解释中，即使模型的可解释性有缺失，算法开发者也仍然有义务对其可能存在的风险进行解释，进而使用户对其建立合理的预期。

其次，国家安全也是重要的考量视角。2020 年 8 月，商务部、科技部调整发布了《中国禁止出口限制出口技术目录》（以下简称《技术目录》）。在本次调整中，在“计算机处理技术”（编号：056101X）项新增“基于数据分析的个性化信息推送服务技术”这一控制要点，并对其出口进行限制。如若个性化信息推送服务提供者对自身算法进行过度的解释，则对算法进行复现或破解的风险大大增加，《技术目录》对其进行的保护也将失去实质上的意义[41]。《技术目录》对算法的规制，体现出了算法解释与国家安全的张力。最后，算法作为诸多企业的核心竞争力，算法解释权与其商业秘密和竞争优势之保护亦存在张力。

因而笔者认为，算法解释权的边界，在于算法可解释性、商业秘密和国家安全等的动态平衡。过于模糊的算法解释不利于对自动化决策相对人权利的保护，而细节化的算法解释则意味着可能的对商业秘密和国家安全的侵害。对于高风险算法来说，应当首先保护国家安全和用户权益，一定程度上牺牲商业利益；而对于中风险算法来说，对其进行解释时，用户权益则需要一定程度上与商业利益平衡。基于此原则，方可动态确立算法解释权的边界，平衡国家、算法使用者与用户之三方利益保护。

结束语 随着算法应用的不断深化，其风险亦逐渐显露，算法规制已然成为学界关注的重要论题。算法规制的目标为建立一套规制算法权利，预防算法权利异化风险，消除算法权利异化后果的制度体系[10]。而算法解释权因其独特的制度价值，逐渐被立法与实践推崇。本文在讨论了算法解释权的立法沿革、现行法规制后，提出了在全算法开发流治理和分级分类解释框架视角下的算法解释权重构策略，即应当建立全算法开发流的治理视角和分级分类的算法解释框架。然而，权利的内在构造、适用范围和行使程序等具体规则，应当根据人工智能等技术的发展而调整。在技术起步发展阶段，适当限定算法解释权行使的范围需充分考虑促进产业发展，这也是现阶段制度设计不容忽视的价值；当人工智能产业得到充分发展后，则应适度放宽行使条件以充分保护用户的权利。算法解释权不仅是一个法学命题，同时也是一个技术命题。算法解释权的制度配置需要兼顾技术发展与法律规制的动态平衡，也需要依据技术的发展路径进行动态调整，这也是未来“计算法学”“数字法治”与“司法人工智能”等新兴交叉学科需要协同努力的方向。

参 考 文 献

[1] YANOFSKY N S. Towards a Definition of an Algorithm [J]. Journal of Logic and Computation, 2011, 21(2): 253-286.

[2] STEINER C, DIXON W. Automate this: How Algorithms Came to Rule Our World[M]. Portfolio/Penguin, 2012.

[3] DIETERICH W, MENDOZA C, BRENNAN T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity [R]. Northpointe Inc, 2016.

[4] ZELEZNIKOW J. An Australian Perspective on Research and Development Required for the Construction of Applied Legal Decision Support Systems [J]. Artificial Intelligence and Law, 2002, 10(4): 237-260.

[5] MCPEAK A. Disruptive Technology and the Ethical Lawyer [J]. The University of Toledo Law Review, 2018, 50: 457.

[6] PASQUALE F. The Black Box Society: The Secret Algorithm that Control Money and Information[M]. Harvard University Press, 2015.

[7] ZHENG Z H. The Ethical Crisis and Legal Regulation of the Artificial Intelligence Algorithm[J]. Science of Law (Journal of Northwest University of Political Science and Law), 2021, 39(1): 14-26.

[8] LI J. The Construction of the Right of Algorithmic Interpretation in Public Services[J]. Seeking Truth, 2021, 48(3): 110-120.

[9] LV B B. On the Algorithm Explanation Obligation of Personal Information Processors[J]. Modern Law Science, 2021, 43(4): 89-101.

[10] ZHANG L H. Regulation of Algorithms in the Age of Artificial Intelligence [M]. Shanghai: Shanghai People's Publishing House, 2021.

[11] BRENNAN T, DIETERICH W, EHRET B. Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System [J]. Criminal Justice and Behavior, 2009, 36(1): 21-40.

[12] BLOCH-WEHBA H. Access to Algorithms[J]. Fordham Law Review, 2019, 88: 1265.

[13] CITRON D K, PASQUALE F. The Scored Society: Due Process for Automated Predictions [J]. Washington Law Review, 2014, 89: 1.

[14] HARARI Y N. 21 Lessons for the 21st Century[M]. Random House, 2018.

[15] ZHOU W. Algorithmic Conspiracy of Antitrust Regulations [J]. Law Science, 2020(1): 40-59.

[16] JIA K. Artificial Intelligence and Algorithm Governance Research[J]. Chinese Public Administration, 2019(1): 17-22.

[17] ZHENG Z H, XU Z X. Legal Regulation and Judicial Review of Algorithmic Discrimination in the Age of Big Data: Take Legal Practice in the U. S. as an Example[J]. Journal of Comparative Law, 2019(4): 111-122.

[18] XIE Z S. Regulating Algorithmic Decision: Focusing on the Right to Explanation of Algorithm[J]. Modern Law Science, 2020, 42(1): 179-193.

[19] ZHANG L H. Research on Algorithmic Interpretation Power of Business Automation Decision-making[J]. Science of Law (Journal of Northwest University of Political Science and Law), 2018, 36(3): 65-74.

[20] JIA Z F. The Right of Algorithm Interpretation is not a Legal

Right—Comment on Article 25 of Personal Information Protection Law (Draft) [J]. Electronics Intellectual Property, 2020 (12): 49-61.

[21] SHAO G S, HUANG Q. Algorithmic Harms and the Right to Explanation[J]. Chinese Journal of Journalism & Communication, 2019, 41(12): 27-43.

[22] WACHTER S, MITTELSTADT B, FLORIDI L. Why a Right to Explanation of Automated Decision-making Does not Exist in the General Data Protection Regulation [J]. International Data Privacy Law, 2017, 7(2): 76-99.

[23] XU K. Taming Algorithms: Historical Evolution and Contemporary System of Algorithm Governance [J]. ECUPL Journal, 2022, 25(1): 99-113.

[24] ZHANG L H. The Iteration and Innovation of Algorithm Regulation[J]. Legal Forum, 2019, 34(2): 16-26.

[25] ZHANG E D. Background, Logic and Structure of the Right to Explanation of Algorithmic Decision-making in the Age of Big Data[J]. Legal Forum, 2019, 34(4): 152-160.

[26] AMERSHI S, BEGEL A, BIRD C, et al. Software Engineering for Machine Learning: A Case Study [C] // 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, 2019: 291-300.

[27] Microsoft. The Team Data Science Process [EB/OL]. (2022-03-03) [2022-04-26]. https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview.

[28] FAYYAD U, PIATETSKY-SHAPIRO G, SMYTH P. The KDD Process for Extracting Useful Knowledge from Volumes of Data [J]. Communications of the ACM, 1996, 39(11): 27-34.

[29] WIRTH R, HIPP J. CRISP-DM: Towards a Standard Process Model for Data Mining [C] // Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. 2000: 29-40.

[30] PAN S J, YANG Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[31] HOGAN B. The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online [J]. Bulletin of Science, Technology & Society, 2010, 30(6): 377-386.

[32] CONG Y N, WANG Z Y, ZHU J Q. Insights into Dataset and Algorithm Related Problems in Artificial Intelligence for Law [J]. Computer Science, 2022, 49(4): 74-79.

[33] ISARAN E T. When an Algorithm Helps Send You to Prison [EB/OL]. (2017-10-26) [2022-04-26]. https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

[34] ZUO W M. Will the Era of AI Judges Come—Based on the Comparison and Outlook of Judicial Artificial Intelligence Between China and Foreign Countries [J]. Tribune of Political Science and Law, 2021, 39(5): 3-13.

[35] SHEN W X. On the Construction and Systematization of the Personal Information Right [J]. Journal of Comparative Law, 2021(5): 1-13.

[36] WEN Y. The Nature and Prospect of Algorithmic Rights—The Theoretical Separation and Functional Compatibility Based on Algorithmic Rights and Personal Information Rights[J]. Journal of Huazhong University of Science and Technology (Social Science Edition), 2022, 36(1): 54-63.

[37] RAVEN B H. Social influence and power[R]. California University Los Angeles, 1964.

[38] ZHANG L H. Function and Realization of the Algorithm Interpretation Rights in Business Automated Decisions[J]. Journal of Soochow University (Philosophy & Social Science Edition), 2020, 41(2): 51-60.

[39] GUNNING D, STEFIK M, CHOI J, et al. XAI—Explainable Artificial Intelligence [J]. Science Robotics, 2019, 4(37): eaay7120.

[40] VLADECK D C. Machines without Principals: Liability Rules and Artificial Intelligence [J]. Washington Law Review, 2014, 89(1): 117.

[41] SU Y. An Interpretation and Specification of the Obligations of Optimizing the Explainability and Transparency of Algorithm [J]. Science of Law (Journal of Northwest University of Political Science and Law), 2022, 40(1): 133-141.

CONG Yingnan, born in 1985, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include big data on business and law, artificial intelligence, blockchain, Fin-tech, Reg-tech and complex system.

ZHU Jinqing, born in 1984, postgraduate, engineer, is a member of China Computer Federation. His main research interests include database systems, content data analysis, artificial intelligence and knowledge graphs.

(责任编辑:柯颖)