中外法学 Peking University Law Journal
Vol.31, No.6(2019)pp.1425－1445

算法解释权与算法治理路径研究

张 欣*

摘 要 面对日益紧迫的算法治理需求,算法解释权被提出,对用户和相关个体的自治性加以尊重,为用户和相关个体的技术性正当程序权利奠定行使基础、避免算法危害成本外化和弥散。欧盟《通用数据保护条例》在立法层面构建了限制和弱化版本的算法解释权,但通过数据主体权利和数据保护影响评估制度在法律实施层面对其加以补强。但其仍然存在制度构造不足、语句模糊不清、适用范围有限等问题。在构建本土化的算法解释权时,应当明晰算法解释权在算法治理中的地位和功用,厘清其行使要件和核心内容,以社会嵌入性和应用领域为基准探索精准化、场景化的衡量机制,打造内外兼具的技术协同治理机制。

关 键 词 算法治理 算法解释权 算法影响评估

引 言

作为数字经济时代的“基础语言”,算法的效能不仅直接决定着决策效率、决策质量,还日益对公私主体产生直接和间接的多维度影响。“算法经济”和“算法社会”的兴起使得算法的研发、投资和应用成为了科技领域的新型增长点。但伴随着算法应用领域的不断扩展和延伸,算法适用危机层出不穷,算法治理体系亟待健全和完善。为保证算法的公平性和透明性,各国政策制定者纷纷探索具有适应国情的、智慧化的治理框架。但无论是以新型算法权利为基础展开算法治理的欧盟,还是依托独立监管和外部审计进行算法治理的美国,都已不约而同地开始关注算法可解释性的立法。我国的算法治理体系也在形成发展之中,算法可解释性以及算法解释权无疑会成为制度设计的首要难点。鉴于这一主题在算法治理理论和前沿实践中的重要

* 对外经济贸易大学法学院副教授。本文受到国家社会科学基金青年项目(项目编号:17CFX058)和对外经济贸易大学中央高校基本科研业务费专项资金(项目编号:CXTD10－04)资助。

意义，本文以算法解释权为研究对象，首先从规范性视角分析其被提出并获得法律和技术共识的三项理据。其次从具体制度构造切入，以欧盟《通用数据保护条例》（以下简称“GDPR”）为基础样本，结合第 29 条工作组指南，系统深入地分析代表性立法的运作机理、策略逻辑和得失利弊。通过分析欧盟在算法解释权立法过程中的重要经验和启示，结合我国算法应用和治理现状，本文对如何构建行之有效的算法解释权提出四项对策建议。

一、算法解释权成为算法治理制度核心的三项理据

基于社会建构与文化情境，商业和市场逻辑下的算法常体现为一系列嵌入具体场景的自动化决策。其之所以得到广泛应用，主要原因之一在于自动化决策的绩效表现常优于人类决策。〔1〕由于规避了决策噪音的存在，算法决策的稳定性、准确性、效率性等优势常常更为突出。〔2〕人类决策不仅在性能层面无法胜出算法，就决策过程而言，人类决策亦非完全透明。虽然“算法黑箱”的存在常引发算法焦虑，但人脑决策也是黑箱。人类决策时常依靠“直觉”这种难以量化的因素，即使是以专业、客观、公平为内核的司法决策，法官决策的影响因素也一直呈现复杂性。〔3〕因此在讨论如何建构行之有效的算法解释权之前，首先需要厘清当人类决策过程尚无法做到全知全悉时，为何各国立法者纷纷要求算法决策具备可解释性，并将其置于算法治理议程的首要位置？本节将从算法技术伦理、技术性正当程序以及算法运行危害成本内化三个层面逐一探讨。

（一）算法解释权对用户和相关个体的自治性以尊重

算法解释权对于用户和相关个体具有独立价值。虽然在万物趋于高度量化的时代其价值难以被数字化地呈现和衡量，但在算法社会中，算法解释权是保障和尊重个体自治性的首道屏障，是机器学习和数据科学领域伦理规范的核心要素，被视为算法时代对抗数据个体的主体性和自治性沦陷和丧失的“内在之善”。〔4〕汤姆·泰勒曾指出，人们自愿遵守法律并非出于恐惧或者自利，实则出于对合法性法律和机构的尊重。〔5〕以中立、无偏、诚实、公正、文明以及

〔1〕 参见（美）纳特·西尔弗：《信号与噪声》，胡晓姣、张新、朱辰辰译，中信出版社 2013 年版，第 125 页。

〔2〕 参见（美）丹尼尔·卡内曼、安德鲁·罗森菲尔德等，“决策的隐形赋税：噪声”，《哈佛商业评论》中文版 2016 年 10 月 27 日，https://www.hbrchina.org/2016－12－09/4780.html，最后访问日期：2019 年 11 月 19 日。

〔3〕 See Richard Posner, *How Judges Think*, Cambridge: Harvard University Press, 2010, pp. 19－57, 125－158.

〔4〕 参见陈璞：“论网络法权构建中的主体性原则”，《中国法学》2018 年第 3 期，第 71－88 页；Andrew Selbst, Solon Barocas, “The Intuitive Appeal of Explainable Machines”, *Fordham Law Review*, Vol. 87, 2018, pp. 1118－1119.

〔5〕 Tom Tyler, *Why People Obey the Law*, Princeton: Princeton University Press, Revised ed., 2006, pp. 3－7.

尊重公民权利为内核的程序本身就具有独立性价值，会对个体行为和感知产生积极影响。〔6〕由此检视当下高速运转、广泛适用、影响深远的算法决策，可以发现当人类争先授权机器做出可能无法逆转的公私决策时，赋予用户或者相关个体获得解释的权利不仅具有工具层面的必要性，更具有规范层面的正当性，是技术时代尊重个体主体性、自治性和人格尊严的基本要求。

从工具视角而言，算法解释权有助于增加人们的算法信任，提升个体接受算法决策的意愿。人类技术采纳行为规律表明，社会面对失去掌控、无从理解、难以救济的新兴技术常会出现焦虑、恐惧甚至厌恶状态。〔7〕伯克利·戴佛斯特等人通过实验发现了“算法厌恶”的认知现象。该现象表明尽管人们已经意识到算法与人类决策相比在预测和评估等领域具有稳定性，但当算法在一项预测性决策中出现与人类决策相同概率的失误时，人们对算法会更快失去信心，并表现出算法厌恶的行为模式。〔8〕不过后续研究表明，虽然人们对算法的预期和信任不足，只要增加人们对算法决策过程的掌控感，算法厌恶情绪会得到明显缓解。在该项实验中，当赋予被试者对算法决策在一定区间范围内调整和修正的权利时，受试者就已表明更有兴趣继续使用算法。〔9〕由此可见，当参与者享有自治性空间和控制能力时，算法信任会有所提升。在价值层面，算法解释权体现了技术社会对人性化的关照和尊重，体现了现代社会应当具有“道德可理解性”〔10〕的诉求，蕴含了算法决策应当符合公平、公正、无偏的内涵和治理目标。更为深层的意义上，当社会空间被信息化，人们的过去、现在和未来都被简化为信息存在时，算法解释权是让人们接受计算化趋势和数字化治理的一个关键联结。〔11〕正如劳伦斯·索勒姆指出的，提供合理、充分的解释是撬动自动化与人格化互动关系的重要节点。〔12〕

（二）算法解释权为用户和相关个体的技术性正当程序权利奠定行使基础

与人类决策相比，算法决策存在本质不同。对于受到决策影响的用户和相关个体而言，适用于传统决策时代的知情权、参与权、异议权和救济权面临诸多挑战。首先，算法模型的存在

---

〔6〕 See Tom Tyler, “What is Procedural Justice? : Criteria Used by Citizens to Assess the Fairness of Procedures”, *Law and Society Review*, Vol. 22, No.1, 1988, pp. 103－132.

〔7〕 See Christain Pieter Hoffmann, Christoph Lutz, Miriam Meckel, “Digital Natives or Digital Immigrants? The Impact of User Characteristics on Online Trust”, *Journal of Management Information Systems*, Vol. 31, No. 3, 2014, pp. 138－171.

〔8〕 Berkeley Dietvorst, Joseph Simmons, Cade Massey, “Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err”, *Journal of Experimental Psychology: General*, Vol. 1, No.1, 2014, pp. 1, 10.

〔9〕 Berkeley Dietvorst, Joseph P. Simmons, Cade Massey, “Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them”, *Management Science*, Vol. 64, No. 3, 2016, p. 1161.

〔10〕 See Andrew Selbst, supra note 4, p. 1118.

〔11〕 See Meg Leta Jones, “The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood”, *Social Studies of Science*, Vol. 47, No. 2, 2017, pp. 216－239.

〔12〕 See Lawrence Solum, “Legal Personhood for Artificial Intelligences”, *North Carolina Law Review*, Vol. 70, No. 4, 1992, pp. 1231－1287.

和运行常处于保密状态，使得相关公众无从知晓其切身利益可能受到机器决策的干预和影响。其次，即使算法运行的事实已被知晓，但控制者常藉由商业秘密或者知识产权为合法理由拒绝披露实际操作原理。再次，即使算法底层代码和运行原理被全面披露，技术性知识鸿沟导致监管者无从问责，数据主体难以有效行使参与权、异议权和救济权。〔13〕面对算法决策带来的诸多挑战，算法解释权成为监管者和数据主体逃脱算法操纵并重获控制的关键一环。2016 年威斯康星州最高法院审理的卢米斯案件以及一系列针对 COMPAS 软件的争论就可深度展现前述三项挑战，以及算法解释权在辅助个体行使正当权利时的重要功用。〔14〕

在卢米斯案件中，被告通过正式程序获知其将接受 COMPAS 软件提供的累犯风险评估，但浮于形式、流于表面的“知情权”不仅限制了被告的抗辩权，算法决策的复杂性、技术性也限制了法院的理解和判断。该案刑事诉讼被告卢米斯主张基于群体行为数据的累犯风险评估存在准确性和科学性风险，侵犯了被告人享有的单独并在准确信息基础上接受审判的权利。卢米斯进一步指出，他虽处于反驳 COMPAS 风险评估的最佳位置，但却无法仅基于开发者在预测报告中提供的有限信息进行质疑，除非他能够知晓影响分数的因素被赋予的权重以及风险评分的运行逻辑。〔15〕面对这一诉求，开发者 Northpointe 称卢米斯所需获知的信息具有专有性，无法向卢米斯以及一审法院公开其评分决策所涉因素的具体权重，但 COMPAS 生成的评估分数具有科学性和准确性。虽然这一观点得到了纽约州刑事司法部门研究报告的支持，〔16〕但质疑的意见此起彼伏。〔17〕面对众说纷纭的判断，审理法院虽然尽量克制和专业，但正如雪莉·阿布拉罕森法官在协同意见中所指出的那样，法院“缺乏对 COMPAS 的理解是一个重大问题”，而“法院因此需要得到帮助”。〔18〕她进一步指出，法院考虑 COMPAS 是否符合正当程序的前提是审判记录中已经提供了有意义的推理过程来阐述该工具的相关性、优势以及弱点。〔19〕量刑法院为国家、被告以及公众提供具有透明性和可理解的解释不仅为上诉法院审查提供了基础，还会增加一审法院对新型决策辅助工具的认知。〔20〕因此，虽然威斯康星

〔13〕 See Andrew Selbst, supra note 4, pp. 1093－1094.

〔14〕 Loomis v. Wisconsin, 881 N.W.2d 749 (2016). COMPAS 全称为提供替代性制裁矫正犯罪管理画像，是运用算法依据累犯和犯罪职业特征相关行为和心理结构设计的为一般和暴力累犯以及审前不当行为的风险预测评估软件。

〔15〕 Ibid.

〔16〕 Sharon Lansing, “ New York State COMPAS－Probation Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy”, Division of Criminal Justice Services Office of Justice Research and Performance, 2012, http://www.northpointeinc.com/downloads/research/DCJS_OPCA_COMPAS_Probation_Validity.pdf, last visited September 21, 2019.

〔17〕 Jeff Larson et al, “How We Analyzed the COMPAS Recidivism Algorithm”, 2016, https://www.propublica.org/article/how－we－analyzed－the－compas－recidivism－algorithm, last visited September 21, 2019.

〔18〕 See Loomis v. Wisconsin, supra note 14.

〔19〕 See Jeff Larson et al, supra note 17.

〔20〕 See Loomis v. Wisconsin, supra note 14.

州最高法院在判决中指出，在限制和谨慎使用的建议被充分考虑的情况下，一审法院在量刑中使用算法风险评估软件并未违反被告的正当程序权利，但仍然应当在未来审理过程中对风险评估算法的使用提供充足的程序性保障措施，包括在预测报告中明确指出评估的优势、相关性以及不足。〔21〕

由此可见，要解决这一难题，从贴合技术架构规律的视角出发，使相关主体在算法决策全过程获得知情权、参与权和异议权需要打开闭环决策回路，切实赋予主体知晓并理解算法运行逻辑的权利，〔22〕进而为公众践行正当程序权利提供行动基础。综上可知，理解算法运作的机理是对可能过大的算法权力予以抵消和施以监督的重要前提，是用户行使技术性正当权利的先决基础，是避免算法决策武断和恣意，保证算法决策可信、正当和理性的重要约束机制。

(三)算法解释权有助于避免算法运行危害成本的外化和弥散

从立法、执法到司法，从资源分配、战略部署到责任认定，一系列公共和商业决策正由算法辅助或者替代执行，日益显著地发挥着系统性和累积性效果。这些算法通过对公民数字身份加以重构，创造了数字化的阶层和类别。在类别重构的过程中，算法决策的系统性和反复性使得在某一领域遭遇算法歧视的主体被结构性锁定，导致其未来的发展和机会受到系统性不利限制。〔23〕但面对算法运行的危害成本，目前却尚无有效的治理措施将其内部化。相反，由于算法决策建立在数据和信息基础上，面对适用算法造成的不良影响，控制者常以其不具有侵权或者故意的主观意图为借口，借助技术复杂性和不可解释性转移算法责任。〔24〕这种将算法责任承担藏匿于技术盾牌之后，希冀逃避算法干预义务、避免承担不利算法责任的做法已经客观上成为算法治理的巨大难题。面对这一挑战，监管者需要在清晰知晓算法运作逻辑的基础上，更新创制出算法责任规则。目前这一制度的缺位给算法控制者推卸法律责任、放任不良运行成本的外化提供了监管套利的空间。

例如，前哈佛数据隐私实验室主任拉坦娅·斯威妮怀疑非洲裔美国人受到了美国互联网在线搜索和广告服务的歧视性对待。她发现在谷歌搜索引擎输入经典非洲裔美国人姓名后，搜索结果的自动配比与刑事犯罪、逮捕记录等负面信息呈现显著相关性。对于这一指控，谷歌及相关网站强烈否认，并以不可解释性这一技术借口进行反驳。〔25〕谷歌指出，搜索引擎算法的主要设计因素是点击率最大化。而人们具体点击的原因无从知晓且无法解释。其算法更多

〔21〕 See Loomis v. Wisconsin, supra note 14.

〔22〕 See Danielle Keats Citron, "Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age", *California Law Review*, Vol. 80, No. 2, 2007, pp. 241－295.

〔23〕 Jack Balkin, "The Three Laws of Robotics in the Age of Big Data", *Ohio State Law Journal*, Vol. 78, No. 5, 2017, pp. 1235, 1237, 1240.

〔24〕 Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge: Harvard University Press, 2015, pp. 39－40.

〔25〕 Ibid.

体现为与投票机类似的中性功能，无法对用户搜索或者表达行为施以影响。〔26〕在无从知晓或系统理解一项算法设计的代码、架构、数据以及运行逻辑时，人们几乎无法对此类反驳进行有效的挑战、回应或裁决。〔27〕故清晰认知和了解算法运行逻辑是判定算法行为主体责任的客观需要。算法解释权要求算法设计者和使用者对模型设计目的、历史以及特定决策原因做出具有意义的逻辑性阐释，不仅可以在一定程度上防止算法控制者逃避责任承担的问题，还可以协助确定算法影响的消极程度从而为监管干预提供介入指引和制度基础。〔28〕

综上所述，算法解释权以及算法可解释性成为了算法归责的重要依据。即使立法者无意为算法决策单独创建一套归责规则，为避免算法运行危害成本的外化，防止社会不公的弥散，立法者也需要建立基础性、前置性的规则作为规制基础。〔29〕确保算法设计具有可追溯性、决策具有可解释性是规则建立的基石。

二、算法可解释性立法样本剖析：GDPR 的启示与镜鉴

由于算法可解释性蕴含上述规范价值，目前已经获得多国立法者的广泛共识。〔30〕纵览全球有关算法治理的相关立法，欧盟立法是值得深入分析和关注的经典样本。通过在 GDPR 中赋予数据主体一系列新型算法权利，欧盟强化了个体对自动化决策的控制权和影响力，构建了以个体赋权为核心路径的算法治理规则体系。为避免算法运行危害成本的外化和弥散，在个性化的算法解释权之外，立法者还通过敦促控制者建立内部问责制度，践行数据保护影响评估建立起持续性、动态性和系统性的系统问责路径。〔31〕本节通过对 GDPR 整部立法进行系统分析，结合第 29 条工作组发布的指南（以下简称“AP29 指南”）〔32〕深入剖析欧盟的算法治理思路。以算法解释权为核心检视对象，系统阐释其构建机理和得失利弊。本节提出，GDPR 虽然在背景引言中构建了限制和弱化版本的算法解释权，但通过知情权、访问权等数据主体权利结合数据保护影响评估制度，欧盟在法律实施层面建立了组合和强化版本的解释权框架。〔33〕

〔26〕 See Frank Pasquale, supra note 24, pp. 39－40.

〔27〕 See Frank Pasquale, supra note 24, pp. 39－40.

〔28〕 See Jack Balkin, supra note 23, pp. 1217－1241.

〔29〕 See Jack Balkin, supra note 23, pp. 1232－1240.

〔30〕 See Government of Canada, “Directive on Automated Decision－Making”, 2019, https://www.tbs－sct.gc.ca/pol/doc－eng.aspx? id＝32592, last visited October 15, 2019.

〔31〕 Margot Kaminski, “The Right to Explanation, Explained”, *Berkeley Technology Law Journal*, Vol. 34, No. 1, 2019, pp. 201－208.

〔32〕 Article 29 Data Protection Working Party, “Guidelines on Automated Individual Decision－making and Profiling for the Purposes of Regulation 2016/679”, February 6, 2018.

〔33〕 本部分观点受到 Bryan Casey, Ashkon Farhangi, Roland Vogl, “Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise”, Vol. 34, No. 1, *Berkeley Technology Law Journal*, 2019, pp. 145－187 的启发。

(一)算法治理的关键条款:GDPR 第 22 条反对仅基于自动化个人决策的权利

GDPR 第 22 条是整部立法中最为直接地对自动化决策以及识别分析进行规定的核心条款。该条文以赋予数据主体不适用仅基于自动化个人决策约束的反对权为基础保护机制,以例外情形下要求数据控制者采取适当保障措施为强化保护机制,结合所涉数据的敏感程度,打造了保护数据主体自主性的多层次联动体系。条文中的核心概念"仅基于自动化的数据处理"需引起注意。这是指一项自动化决策制定过程中完全没有人工干预。相反,若在决策过程中有实质意义上的人工干预,则不属于本条规定的对象范围。但为避免数据控制者投机性地规避条款,AP29 指南中表明人工干预应当具有实质性监督意义,应当由具有权限和能力对决策结果进行改变和影响的人执行。〔34〕

进一步而言,该条虽然以限制企业对数据主体使用具有法律影响或者类似重要影响的完全自动化决策为目的,但考虑到实践中场景和利益的多元化,因而规定了(a)合同履行、(b)经欧盟或者成员国法律授权以及(c)基于数据主体明确同意这三种可以对数据主体适用自动化决策的例外情形。虽然该三项例外情形可以作为自动化处理的合法和适当基础,但考虑到实践中自动化决策所涉双方力量对比悬殊的现状,第 22 条在第 3 款提出即使属于(a)(c)两款中规定的例外情形,数据控制者也不应消极放任,而是应当"采取适当的措施"保护数据主体的权利、自由和合法利益。此处的"适当措施"指向数据主体获得与反对权相关的衍生权利,包括获得人工干预、表达其观点以及提出异议的权利。

第 22 条对于完整理解 GDPR 算法治理思路十分重要。鉴于自动化决策的专业性、复杂性,欧盟立法者并未像美国一样直接诉诸于外部问责机制,而是以立法限制的方式赋予数据主体不受仅单纯自动化决策支配的权利。这一条文设计充分考虑了算法问责成本高昂、信息获取不足导致外部监督乏力等实践现状。〔35〕当控制者具有合法正当的适用理由时,欧盟立法者通过深入到自动化决策双方的互动机制之中,在逻辑层面构筑了表达权、异议权和获得人工干预权,与反对权形成联动,搭建起保护个体算法权利的基础体系。考虑到个体享有的三项权利在实施层面可能遇到困难和挑战,GDPR 在背景引言中引入算法解释权概念,确保数据主体在实质层面理解一项自动化决策作出的逻辑和依据,以作为第 22 条数据主体有效行使反对权、表达权、异议权等算法权利的制度基础。后文的分析将进一步表明,背景引言的算法解释权与第 13、14 条赋予数据主体的知情权有机耦合,从实质和形式两个层面系统保障数据主体获知决策相关信息,确保其具有提出异议并获得人工干预的信息能力。

(二)限制和弱化版本的算法解释权:GDPR 背景引言 71

依据第 22 条设定的治理结构,结合上文中算法解释权预期实现的三项价值为分析框架,

〔34〕 See Margot Kaminski, supra note 31, p. 205.

〔35〕 Ansgar Koene, Chris Clifton, Yohko Hatada, Helena Web, Rashida Richardson, "A Governance Framework for Algorithmic Accountability and Transparency", April 2019, p. 1, http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf, last visited November 15, 2019.

可以对 GDPR 背景引言 71 所规定的自动化决策解释权进行梳理和分析。总体而言,欧盟立法者在背景引言中设定了弱化和限制版本的解释权。这一策略体现了立法者在技术发展、商业秘密和知识产权制度等多元利益间寻求微妙平衡,彰显了立法层面的克制和谨慎。就当前技术发展阶段而言,机器学习模型设计仍然需要在模型可解释性与决策准确性之间权衡取舍。〔36〕若冒然实施完整版本的算法解释权可能会损失决策绩效,成为扼杀创新和技术发展的制度障碍,从而对社会效用带来不利影响。但仅依靠弱化和限制版本的算法解释权难以达成三项预期价值。

首先,背景引言算法解释权的适用场景未能突破第 22 条设定的仅基于自动化决策的限制,对于在实践中广泛存在的人机结合决策场景无法适用,故对数据主体的自治性和主体性保障不足。其次,背景引言算法解释权语句规定模糊,制度构造仍存多处缺陷,难以作为用户行使抗辩权的基础。例如,条文中"对评估后作出的决定"这一限定表述表明此处的解释权是存在于事后阶段且面向一项具体决定的。若无其他规范支撑,难以从现有条文中推导出数据主体获得系统性解释以及事前和事中解释的权利。仅聚焦于事后阶段的解释权无法给数据主体妥当全面地提供到底是否应当明确同意接受一项自动化处理的判断,造成权利行使与系统运行的断裂与错位,并未与第 22 条良好衔接。此外,算法解释权的内涵、标准等核心内容规定阙如。背景引言 71 中对于控制者应当提供的解释内容语焉不详,数据主体难以据此主张要求获得有关模型设计原理、变量特征、变量权重等核心解释内容。因此,背景引言 71 设定的解释权版本难以保障数据主体做到实质意义上的知情,无法成为参与、异议以及获得救济等正当程序权利的行使基础。最后,背景引言的算法解释权难以避免算法运行危害成本的外化和弥散。一方面,背景引言 71 以数据个体为基础,限制了数据主体采取集体行动的制度空间,无法有效应对自动化决策在社会层面带来的潜在系统性问题。另一方面,背景引言的算法解释权对设计者探索更具解释性算法模型的激励作用有限。正如准确的产品质量评级制度可以激励生产者做出更好的符合优质评级的产品一样,〔37〕算法解释权制度本可以成为帮助算法设计者提升系统可解释性和透明度的有力激励,但背景引言的算法解释权呈现限制和弱化特征,未能从实质层面为算法开发者在设计阶段提供明确清晰的行动指引,难以实现避免算法危害成本外化和弥散的功用。

(三)补强算法解释权的数据权利条款:以 GDPR 第 13—15 条和第 35 条为例

虽然 GDPR 生效后算法解释权的相关规定一度引发了行业恐慌和担忧,但背景引言不具有单独的法律效力,〔38〕同时立法者受到诸多限制,导致立法层面的解释权趋于弱化。但鉴于

〔36〕 Madalina Fiterau, "Trade—offs in Explanatory Model Learning", March 2012, https://www.ml.cmu.edu/research/dap—papers/dap_fiterau.pdf, last visited October 20, 2019.

〔37〕 Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", *Nature Machine Intelligence*, Vol. 1, 2019, p. 210.

〔38〕 Lilian Edwards, Michael Veale, "Slave to the Algorithm? Why a Right to an Explanation is Probably Not the Remedy You Are Looking For", *Duke Law & Technology Review*, Vol. 16, No. 1, p. 21, 50.

解释权对算法治理的重要意义，欧盟监管机构提出在实施层面打造更为普遍和强化版本的算法解释权策略。〔39〕欧盟执法机构希望借鉴数据生命周期管理的视角，通过 GDPR 第 35 条数据保护影响评估制度与解释权条款协同配合，从而促使自动化决策在整个设计和运行周期内受到监督、评估和审计。〔40〕此外，由于欧盟的算法治理路径被牢固镶嵌在数据治理框架之中，因此虽然背景引言 71 的算法解释权完整性阙如，但 GDPR 第 13－15 条赋予了数据主体知情权、访问权。通过获得这些数据权利，数据主体在一定程度上获得了补强和组合版本的算法解释权：

首先，第 13 条和第 14 条赋予数据主体知情权，在一定程度上弥补了背景引言解释权在事前阶段关照不足的缺陷。如前所述，背景引言的算法解释权仅聚焦于一项自动化决策的事后阶段，而在事前和事中阶段未能对数据主体提供有力的制度支持。第 13 条和第 14 条规定了数据控制者收集信息和获取个人数据时应当履行的事前通知义务，可以为数据主体告知其目的等信息奠定制度基础。具体而言，第 13(1)(c)条和第 14(1)(c)条要求控制者在收集信息时向数据主体提供个人信息处理目的以及处理法律基础的相关信息。第 13(2)(f)条和第 14(2)(g)条规定控制者在获取个人数据时，应当向数据主体提供自动化处理过程中运用的逻辑以及该种数据处理对数据主体的重要性和可能产生的后果。其中，“自动化处理运用逻辑”的信息包括处理过程背后的原理或者决策做出的标准。AP29 指南表明，虽然数据控制者无需总是尝试解释或者全部披露其使用的复杂算法，但其提供的与原理基础相关的信息必须是对数据主体有意义的。依据指南示例，这些信息包括设计时考虑的主要特征、信息来源相关性以及该原理可能导出的任何结论。〔41〕在具体呈现方式上，背景引言 60 以及 AP29 指南要求数据控制者应当以“易见、易懂、易读”的方式提供真实的、切实可能产生的影响类型示例，包括以可视化技术等方式呈现。〔42〕

其次，第 15 条赋予数据主体访问权在一定程度上弥补了背景引言解释权未能覆盖事中阶段的立法缺陷。依据该条，数据主体有权向控制者确认与其相关的个人数据是否正在被处理，以及有权要求访问与其相关的个人数据并获知详细信息。这些信息包括处理目的、相关个人数据的类别、自动决策机制和识别分析过程中运用的逻辑、该种处理对数据主体的重要性以及可能产生的后果等内容。赋予数据主体在“数据处理过程中”的访问权客观上起到了部分行使事中阶段解释权的效果。AP29 指南中指出，如有必要，控制者可以提供支持让专家进一步验证自动化决策的工作过程。这也从专业性方面为数据主体获得更有意义的事中解释提供了指引。但遗憾的是，无论是第 13、14 条规定的知情权，还是第 15 条规定的事中阶段访问权，都限定于第 22 条设定的仅基于自动化决策情形，自动化处理作为决策辅助的情形未能被涵盖其

〔39〕 See Bryan Casey et al, supra note 33, pp. 187－188.

〔40〕 See Bryan Casey et al, supra note 33, pp. 169－176.

〔41〕 See Article 29 Data Protection Working Party, supra note 32, p. 13.

〔42〕 See Article 29 Data Protection Working Party, supra note 32, p. 13.

中。与此同时,无论是事前阶段的知情权还是事中阶段的访问权,均先于一项自动化决策的完成而存在,数据主体由此获得的解释只局限于系统功能,而无法支持数据主体获得更具实质意义的针对一项具体决策的事后解释权。〔43〕虽然理论层面第 13－15 条与背景引言 71 创设的解释权有机组合后可以形成较为完整的解释权,但前述个体权利路径在实践层面面临的诸多限制仍然使得这一方案难以周全保障算法治理的有效性。因此,应当寻找一种平衡和互补的方案,在个体权利路径之外创设撬动数据控制者责任的机制,弥补个体权利路径激励不足的问题。

在这方面,GDPR 第 35 条确立的数据保护影响评估制度被欧盟数据监管机构作为与背景引言算法解释权协同发挥治理效果的补强方案而引发关注。〔44〕该条着眼于数据控制者的隐私风险管理责任意识,有效弥补了个体权利路径的种种局限。首先,第 35 条要求控制者在数据处理前就应进行数据保护影响评估,有助于数据主体在事前阶段获得行使解释权所需的核心信息。有学者指出第 35 条(1)款与第 22(1)款措辞上惊人的相似性并非巧合。两者相互结合会产生强大的协同治理效应。〔45〕与此同时,数据保护影响评估并非仅局限于设计阶段,立法者要求其在循环、连续的基础上使用,体现了动态性、多层次、以数据主体为中心以及闭环治理的特性。其次,第 35 条突破了第 22 条、13－15 条以及背景引言 71 对于自动化决策类型的局限性规定,弥补了 GDPR 在人机结合决策场景的立法空白。依据条文可知,其不仅适用于第 22 条(1)款规定的单独自动化决策的情形,还包括“非完全自动化”的决策情形。〔46〕由此可见,数据保护影响评估制度要求控制者在处理前开始实施,贯穿设计、运行、部署、结束等全周期,是一种动态性的责任机制。为切实保障数据保护影响评估的实施效果,GDPR 在罚则部分规定,未执行第 35 条数据保护影响评估义务的控制者将会被处以一千万欧元或者上一财务年度全球营业总额 2%的罚款。这一机制将算法管理成本转移到控制者,形成了风险管理的闭环和危害风险外部化的转移。〔47〕

综上可知,考虑到算法治理的复杂性和动态性,GDPR 在立法层面采取审慎和克制的立法思路,于背景引言 71 中设定了限制和弱化版本的解释权。但在具体实施层面,知情权、访问权、数据保护影响评估制度分别在事前、事中和事后阶段补强了背景引言的解释权框架,通过对数据主体和数据控制者“双管齐下”,实现了数据治理和算法治理的系统联结。表一通过梳理构建算法解释权的关键要素对算法解释权的立法思路进行系统总结。由表一可见,组合版

〔43〕 Sandra Wachter, Brent Mittelstadt, Luciano Floridi, “Why a Right to Explanation of Automated Decision－Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law*, Vol. 7, No. 2, 2017, pp. 83－84.

〔44〕 See Bryan Casey et al, supra note 33, pp. 169－176.

〔45〕 See Bryan Casey et al, supra note 33, p. 179.

〔46〕 See Article 29 Data Protection Working Party, supra note 32, p. 24.

〔47〕 参见程莹:“风险管理模式下的数据保护影响评估制度”,《网络与信息安全学报》2018 年第 8 期,第 68－69 页。

本的制度设计不仅作为数据主体行使知情、参与、异议和补救的行动基础，还作为贯穿算法设计、测试、部署、完善等多环节的重要监督机制而存在。

表 1　GDPR 背景引言算法解释权与补强条文组合机理

构建算法解释权的关键要素		GDPR 背景引言 71 设定的算法解释权	GDPR 相关条文对算法解释权的补强规定
1.自动化决策的类型	单独自动化决策	有规定	第 22 条
	人机辅助情形决策	无规定	第 35 条(3)款(a)项
2. 所涉自动化决策具体阶段	事前	无规定	第 13 条(1)(c)＋(2)(f) 第 14 条(1)(c)＋(2)(g) 第 35 条
	事中	无规定	第 15 条(1)(h) 第 35 条
	事后	有规定	第 35 条
3. 解释内容	收集数据的信息	有规定	第 13 条、第 14 条
	运行逻辑(数据处理背后的推理知识)	无规定	第 13 条(2)(f) 第 14 条(2)(g) 第 15 条(1)(h)
	系统功能	无规定	第 35 条(7)款
	设计目的	无规定	第 13 条(1)(c) 第 14 条(1)(c)
	具体处理决定	有规定	
	处理重要性	无规定	第 13 条(2)(f) 第 14 条(2)(g)
	设想用途 预期后果	无规定	第 13 条(2)(f) 第 14 条(2)(g)

组合版本的算法解释权体现了立法者在商业秘密保护、知识产权保护、算法问责等多元目标间寻求微妙平衡，通过建立数据主体的权利行使机制和数据控制者的保护责任机制打造算法治理的二元协同框架。〔48〕但组合版本的算法解释权仍存在诸多局限。首先，各条文在综合适用时边界不清，数据主体难以清晰知晓到底应当如何策略性地对上述条文进行组合适用从而为自身获得抗辩权能奠定基础。其次，数据质量控制虽然对保障算法可靠性具有重要作

〔48〕 See Bryan Casey et al, supra note 33, pp. 143, 152, 179.

用,但仅对控制、防范和解决算法模型依赖的数据偏误具有较好效果,对于模型设计偏误难以全面覆盖。最后,GDPR 整部立法中未能明晰控制者提供算法解释的衡量标准。这一核心内容的缺失可能导致算法解释流于形式。正如有批评指出的,向用户提供流于形式的算法解释百害而无一利,因为其可能为控制者作恶提供"虚假的掩饰"。〔49〕但无论如何,作为算法可解释性落实在立法层面的经典样本,GDPR 仍然值得各国立法者深入探究并有选择地予以借鉴。

三、算法解释权构建的本土化建议

在进一步深入探讨算法解释权和算法治理的完善方案之时,还需清晰认知算法治理的根本原则和所应遵循的根本规律。依据不断提升的算力和海量扩充的数据,算法技术被内嵌于平台经济和技术社会的多元场景之中,不断更新和扩充着自己的角色。在日益复杂的算法应用场景中,其扮演着相互交织的三重角色。首先,算法依据特定计算模型将数据转化为可预期结果,作为高效、精准的编码程序体现为纯粹的工具价值,是一种技术和物质存在。其次,算法逐步超越程式推理系统,作为基础设施嵌入到平台经济和技术社会之中,将时间、空间、关系和话语等多种社会要素相互连接,扮演辅助人类在特定场景下进行资源再分配的中介性角色。〔50〕再次,端赖于自身计算精准、运转高效、超出人类计算能力的特质,算法在众多场合成为人类决策的"代理人"和"接管人",〔51〕开始担当自主决策的主体,并成为"社会权力"的构成部分。在第二和第三层面上,作为"社会权力"的算法通过分类、排序、过滤、推荐、预测、评估等技术组合,直接塑造人们被对待的方式和预期机会,对社会关系和社会秩序予以重塑。〔52〕从纯粹的计算工具到中介性基础设施,再到人类决策的"代理人"和"接管人",算法的社会嵌入性不断增加,在人类构建的社会秩序中完成了执行性角色到代理人角色的转换。因此,在谈论算法治理时,绝不应将其视为社会生态之外单纯的技术存在,而须关注其在代码基础上不断重塑社会秩序的决策能力。〔53〕正如有学者指出的,算法通过不断接管人类决策已经成为实质意义上社会秩序中权力部署的一部分。〔54〕因此,在对社会、文化和各种制度结构不断发挥嵌入

〔49〕 See Toon Calders, Indré Žliobait, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures?", *Discrimination and Privacy in the Information Society*, Vol. 3, 2013, pp. 43－57.

〔50〕 参见孙萍:"如何理解算法的物质属性——基于平台经济和数字劳动的物质性研究",《科学与社会》2019 年第 3 期,第 50、52－53 页。

〔51〕 David Beer, "The Social Power of Algorithms", *Information, Communication and Society*, Vol. 20, No. 1, 2016, p. 5.

〔52〕 Ibid., p. 2.

〔53〕 See David Beer, supra note 51, p. 4.

〔54〕 See David Beer, supra note 51, p. 11.

性影响时，仅要求算法保证计算准确和运转高效已经无法与其作为“社会权力”的角色相匹配。

这其中蕴含的紧张关系可借由韦伯对形式理性与实质理性的区分来阐明。在韦伯的语境中，形式理性体现为“超越个别、具体的，以普遍、抽象的规则和可计算的程序为依归，在追求目标的过程中作出合理的安排”。〔55〕例如“纯粹从技术上看，货币是最为完善的经济计算手段，即在经济行为取向的形式上是最合理的手段。”〔56〕但工具和形式本身有效并不必然服务于“正确的目的”，并不一定与道德或者其他形式的价值观所符合。〔57〕因此，韦伯指出“货币计算形式上的合理性本身丝毫不说明事务的实质分配的方式……形式上的合理性只有与收入分配方式相结合，才能说明物质供应的方式。”〔58〕这一规律亦可投射到算法治理领域。从根本上而言，算法的治理应当实现形式理性与实质理性、工具理性与价值理性的平衡，一方面要实现算法作为技术工具的功利最大化，另一方面要实现算法作为中介者和决策代理者在价值承载上的最大化。换句话说，算法治理架构的设定不仅需要促进算法实现基于计算、竞争、效率、客观等优势的形式理性，还需要打开算法作为社会权力的维度和视角，确保其决策过程和决策结果是公平、可信、可责的，将其视为权力理性的一部分，〔59〕秉持这一原则有助于实现算法治理的有序高效，兼顾算法治理的正当性及其对经济社会的总体影响。

从权力约束机理来看，首先需要在现有规则体系中为人们知晓、理解、选择和干预算法决策开辟入口，而算法解释权正是搭建入口的重要方案之一。下文将结合算法解释权蕴含的三项规范价值和欧盟立法样本的经验，对算法解释权构建的本土化路径提出四项建议。

（一）明晰算法解释权在算法治理中的地位和构建策略

如前所述，算法解释不仅是构建算法规制框架的关键节点，也是避免算法决策恣意武断、有效约束算法决策的前提。〔60〕就当下技术发展阶段而言，算法解释权对构建有效的算法治理制度还具有两项特殊功用。依据算法开发的客观情况，算法模型的不可解释性主要由两种原因导致：其一在于为完成具有挑战性的决策任务，算法模型在设计上具有复杂性，从而一定程度上牺牲了可解释性。但反观人类技术的发展规律，当立法设计更加主动、动态地响应技术发展时，技术创新和应用才有可能带来更为积极的整体社会效用。因此虽然技术行业常以客

〔55〕 张德胜、金耀基等：“论中庸理性：工具理性、价值理性和沟通理性之外”，《社会学研究》2001年第2期，第35页。

〔56〕 王俊敏：“韦伯的理性‘进步’及其意义问题”，《社会学研究》2011年第2期，第112页。

〔57〕 参见（德）马克思·韦伯：《韦伯作品集（V）：中国的宗教：宗教与世界》，康东、简惠美译，广西师范大学出版社2004年版，第458－460页；（德）马克思·韦伯：《韦伯作品集（Ⅶ）：社会学的基本概念》，顾忠华译，广西师范大学出版社2005年版，第32页译注。

〔58〕 （德）马克思·韦伯：《经济与社会》（上卷），林荣远译，商务印书馆1997年版，第129页。

〔59〕 See David Beer, supra note 51, pp. 1－13.

〔60〕 Frank Pasquale, “Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society”, *Ohio State Law Journal*, Vol. 78, No. 5, 2017, p. 1239.

观技术限制作为反对算法解释权立法的有力借口,但算法解释权的相关立法其实可以为技术开发者探索更具透明性和高性能的技术模型提供充足的制度激励,平衡市场收益和社会福利,更大程度地释放技术价值。算法模型不可解释性的另一原因在于开发者的商业逐利本性。〔61〕商业秘密和知识产权制度使得算法开发活动罩上了一层神秘色彩。诸多学者主张算法解释权与商业秘密和知识产权制度存在冲突。〔62〕但深究这一问题的根源可以发现,借助商业秘密与知识产权制度的强力保护,算法开发者不仅在与算法可解释性相关的技术开发上怠于行动,其甚至具有将算法模型不断复杂化,借助复杂性和黑箱性获利的情形。〔63〕因此,在立法层面赋予数据主体获得解释的权利,要求设计者履行算法解释义务,以此为突破点将设计者和控制者纳入到算法归责框架之中,扭转和约束技术创新中资本不当逐利引发的监管困境,如此方能有力保护公众免受算法不当侵害。

明晰了算法解释权在算法治理中的功用之后,需要据此结合实践情况选择适当可行的立法策略,分别在短期、中期、长期设定精细化的规则和配套措施,逐步解决算法可解释性的现实局限。欧盟立法者在与代表商业秘密和知识产权利益的主体的立法博弈中虽然有所妥协,〔64〕仅设定了限制和弱化版本的算法解释权。但在机制设计上,其方案设计逻辑十分清晰。欧盟立法者并未孤立地看待算法解释权,而是将这一权利嵌入到第 22 条代表的算法治理架构之中,通过对数据主体和数据控制者"双管齐下"促使算法解释权与数据主体权利以及数据控制者义务等制度设计相辅相成,实现数据治理和算法治理的系统联结,盘活了算法解释权在整体机制设计中的多重功用。算法解释权实质上是立法者设计的促进算法透明度的个性化措施,能够较好地依据算法应用的多元化场景为个人提供灵活和及时的制度保障。但如前所述,个体路径具有离散性特点,无法对算法系统有效问责,且偏误纠正能力十分有限。因此,立法者在机制设计时,还可考虑将个性化的算法解释权与系统性的算法问责制度有机勾连,将算法解释权作为连接算法影响保护和算法归责的重要连接点,持续、动态并贯穿始终地对算法展开引导和监督。〔65〕通过探寻兼具个体利益和集体行动的治理框架,实现个体化和系统化治理路径的深度融合,更好地发挥协同效果,实现治理目标。

(二)厘清算法解释权的行使要件和核心内容

GDPR 的立法经验启示我们:算法解释权是一种新型权利,当其内核未能厘清、衡量标准

〔61〕 See Cynthia Rudin, supra note 37, p. 209.

〔62〕 See Guido Noto La Diega, "Against the Dehumanisation of Decision-Making-Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information", *Journal of Intellectual Property, Information Technology and E-Commerce Law*, Vol. 10, No. 2, 2018, pp. 1-34.

〔63〕 See Cynthia Rudin, supra note 37, p. 209.

〔64〕 See Sandra Wachter et al, supra note 43, pp. 81-82.

〔65〕 See Margot Kaminski, supra note 31, p. 205.

模糊不清时，可能导致其无法有效实施而只能尘封搁置。因此，应当从权利行使要件入手，对该项权利的内核和外延予以精细化设计，以此探索切实可行的制度框架，为实际应用奠定良好的制度基础。对应于自动化决策系统，算法解释权的制度设计至少需要明确两项内容。

第一，应明确解释权面向的对象是自动化系统还是特定决策，亦或通过精细化的场景设置兼括二者。GDPR 条文中仅提及解释的内容包括决策做出所涉及的逻辑、决策的重要性以及对个体可能产生的预期后果。这种立法方式看似清晰，实则指向模糊，缺乏对实践的清晰指引。于立法者而言，如若围绕系统功能构建解释权，应当明确解释信息至少还应包括一般功能化信息。这些信息包括但不限于系统的需求规范、预定义模型、训练参数、输入数据摘要、运行逻辑、模型测试、训练或者筛选的相关信息等。〔66〕倘若解释权被设计为指向某一具体的自动化决策，则应当对具体决策产生的理由、原则和个体数据情况，例如具体特征和功能加权、特定决策规则、参考或者识别分析组的相关信息提供解释。在特定情形下，还应当对前述系统信息做出周全解释。〔67〕在立法形式上，基于算法决策场景多元化和复杂化趋势，建议在立法中明确解释对象，由相关协会或者专门机构发布解释指南予以精细规定。

第二，应明确数据主体行使解释权的阶段和时机。如前所述，GDPR 中出现了事前、事中、事后阶段衔接断裂导致权利行使未覆盖全过程的问题。实际上不同阶段的解释权各具功用。事前的算法解释权可以让数据主体关注算法系统的设计目的、运行方式和决策逻辑，对算法可能带来的影响有所准备。事后的算法解释权可以让数据主体对特定决策的运行逻辑清晰知晓，为其提出异议获得救济提供重要基础。与此同时，解释权行使时机在内容上与解释对象紧密相关。如果赋予数据主体在一项具体决策做出之前和之中主张解释的权利，匹配的解释对象多为系统功能而非特定决策。相应的，如果赋予数据主体在一项具体决策做出之后行使主张解释的权利，则可以面向系统和特定决策两者同时行使。〔68〕对于复杂的算法架构，系统解释并不一定能够推导出特定决策的解释，因此区分解释权涉入的时机和阶段具有重要意义。例如在自动化信用评分场景中，对客户做出具体决策前，该系统的提供商可以告知并解释该系统运作的一般逻辑、系统设计目的、功能和意义以及设想的具体后果（例如利用信用评分由贷方评估个体的信用状况可能影响贷款的批准以及利率等事项）。〔69〕而在特定决策做出之后，则可以在系统功能解释的基础上，向数据主体针对个人数据情况做出具体解释，例如该数据主体的信用评分是如何产生的、在评分产生

〔66〕 See Margot Kaminski, supra note 31, pp. 214－215.

〔67〕 See Sandra Wachter et al, supra note 43, p. 78.

〔68〕 See Andrew Selbst, Julia Powles, "Meaningful Information and the Right to Explanation", *International Data Privacy Law*, Vol. 7, No. 4, 2017, pp. 240－241.

〔69〕 See Sandra Wachter et al, supra note 43, pp. 78－79.

过程中使用的数据和特征以及其在决策树中的相应权重。〔70〕因此，事后阶段的解释权至关重要，其包含与数据主体有关的具体决策过程的细节。尤其在消除算法歧视和数据主体信息不对称方面，一般性的有关算法系统和功能的事前解释过于模糊并缺乏透明度，无法有效支持数据主体获得关于实际决定的具体解释，并由此获得完整意义的解释权。〔71〕故立法者需要周全考虑算法解释权行使的核心内容，秉持动态发展的视野，根据技术发展客观情况，逐步探索适合技术创新与公共利益的算法解释权构成体系。

(三)以社会嵌入性和应用领域为基准起点探寻算法解释义务衡量标准

精准化、场景化地构建算法解释义务对于解释的准确性和效率性至关重要。与人类决策不同，算法系统不会自动存储形成其决策依据的信息。因此，是否需要提供生成解释的功能是在系统设计阶段就需要解决的资源分配问题，应预先对解释性能予以规划，确保算法系统被设计为精确存储输入、中间步骤和输出信息的模型。〔72〕但解释伴随成本，技术设限可能不利于中小企业发展并挤压竞争。〔73〕因此，为了最大化社会总体效用，立法者应当结合具体应用场景探索最佳制度设计方案，通过明确算法解释内容和衡量标准，确保解释系统忠实于原始模型，同时避免不适当的法律要求影响准确性阈值。GDPR 虽未能周全覆盖具体决策类型，但立法者对完全基于自动化和非完全自动化类型的界分在一定程度上体现了分类治理、精准施策和关注场景的立法思路。深究 GDPR 对决策类型的划分机理，可以发现是否存在具有实质意义的人工干预是立法的首要划分依据。在自动化决策过程中，是否存在“人工干预”这一类型化逻辑反映的是算法的社会嵌入性问题，即算法在一项决策中是仅作为执行人类决策的计算工具、一种单纯的技术和物质存在，还是已经成为人类决策的代理人和接管者。〔74〕但 GDPR 类型划分的局限性在于考虑的基准过于单一，由此构建的算法解释权与现实中算法应用的复杂性存在错位。事实上，就网络治理而言，诸多学者对于如何构建场景化的治理曾提出建议。〔75〕例如，海伦·尼森鲍姆的情境完整性理论提出在社会技术系统中，情境是结构化的社

〔70〕 See Sandra Wachter et al, supra note 43, pp. 78－79.

〔71〕 See Gianclaudio Malgieri, Giovanni Comande, “Why a Right to Legibility of Automated Decision-making Exists in the General Data Protection Regulation?”, *International Data Privacy Law*, Vol. 7, No. 4, 2017, pp. 243－265.

〔72〕 See Finale Doshi-Velez, Mason Kortz, “Accountability of AI Under the Law: The Role of Explanation”, Berkman Klein Center Working Group on Explanation and the Law, *Berkman Klein Center for Internet & Society Working Paper*, https://dash.harvard.edu/handle/1/34372584, last visited October 15, 2019, p. 9.

〔73〕 Ibid., p. 12.

〔74〕 See David Beer, supra note 51, pp. 1－13.

〔75〕 See Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies*, Cambridge: Harvard University Press, 2018, pp. 157－197.

会场景，与行为、角色、关系、权力结构、社会规范、内部价值密切相关。〔76〕情境既可以指代应用的条件，又可以构成特定主体的行动环境。因此，在治理中应当尊重语境，通过具体场景探寻信息规范结构。〔77〕但就算法解释义务衡量标准而言，不同的算法应用场景意味着不尽一致的治理要求，所要考虑的情境完整性和精确性目标随着决策类型、所涉主体、社会规范等因素不尽相同。〔78〕前述理论在具体应用时还需具体结合各国立法实际量体裁衣。

本文认为，在众多影响算法决策的治理因素中，算法的社会嵌入性和应用领域应是设定衡量基准时的首要考虑因素，可以成为我国构建动态性、精准性算法解释义务标准体系的参考。图一以算法社会嵌入性与应用场景为基准，提供了算法解释义务精准化治理的一个示例。以算法社会嵌入性为基准，算法在决策中的角色可以分为纯粹执行者和独立决策者；〔79〕以算法应用领域为基准，算法应用场景可以划分为商业领域和公共事业领域两个类别。两个基准有机结合可以细分为四个主要场景。从场景 1 至场景 4，算法社会嵌入性不断加深，算法在决策过程中的自主性和主导性不断增加，控制者解释义务的衡量标准应当逐步趋向严格，算法透明度应逐步提升，控制者对于确保算法可靠性的措施也应逐步增加。〔80〕

就具体的算法解释义务衡量标准而言，目前主要有易读性标准、反设事实标准和可验证标准。三种标准既有交叉，又有不同。依据 GDPR 和 AP29 指南，可以发现欧盟立法者倾向于采用易读性标准，〔81〕即控制者不仅应当提供自动化决策系统或者特定决策所涉相关信息，还需以一种易于数据主体理解的方式呈现。例如，AP29 指南指出，为了让解释更易接受和理解，算法控制者应当给出具体影响类型示例加以说明。比如为数据主体提供应用程序，展示虚拟司机所具有的危险驾驶习惯是如何在算法决策制定过程中影响汽车保费的。〔82〕但这种标准成本较高，并非所有类型的算法解释都可轻易满足。例如，将自动驾驶汽车传感器所收集的高维数据转化为人脑中视觉输入的树木或者街道标志等对应概念就会带来成本和技术挑战。〔83〕因此，有学者提出了反设事实标准。该标准仅要求数据控制者提供和披露满足反设事实假设相关的变量信息，〔84〕通过回答什么是决策中与事实具有因果影响的重要因素来回

〔76〕 Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Redwood: Stanford University Press, 2010, p. 132.

〔77〕 Ibid., pp. 132－145.

〔78〕 See Helen Nissenbaum, supra note 76, pp. 132－145.

〔79〕 两个坐标代表算法决策的两种情形，交叉之处还包括上文中指出的中介者角色。

〔80〕 See Cynthia Rudin, supra note 37, pp. 206－215.

〔81〕 See Gianclaudio Malgieri et al, supra note 71, p. 244.

〔82〕 See Article 29 Data Protection Working Party, supra note 32, p. 10.

〔83〕 See Finale Doshi-Velez et al, supra note 72, p. 8.

〔84〕 See Sandra Wachter, Brent Mittelstadt, Chris Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR", *Harvard Journal of Law & Technology*, Vo. 31, No. 2, 2017, pp. 1－52.

答某个因素是否决定了结果，从而帮助数据主体获得对特定决策的理解。由于这一标准满足了一定程度的可阐释性和易读性，减轻了对商业秘密的关注，同时在有限度的透明度约束下提供了对特定决策的解释，因此这一进路又被形象地描述为“在不打开黑箱的情况下解释黑箱”。〔85〕区别于前述两种标准，一些技术专家认为鉴于当前技术发展现状，如果可以对特定决策重复验证，也可视为提供了解释，本文将其称为最低限度的可验证标准。例如美国 IEEE 协会提出，在特定技术条件下，只要通过技术手段可以对特定决策加以验证，就在一定程度上认为模型具有可解释性。〔86〕以上三种标准成本收益各异，建议根据算法社会嵌入性和应用场景的影响评价将算法解释义务衡量标准精细化、类型化，促使开发者在设计阶段预先调整。

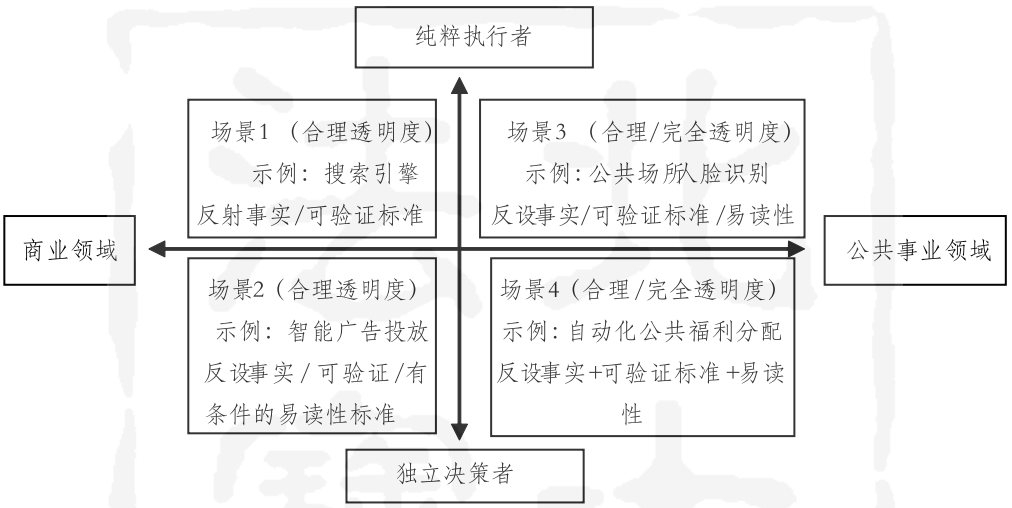

图 1 算法解释义务衡量标准场景化示例

当然，鉴于算法架构复杂性和应用场景多元性，本文提出的划分基准主要是一个有益参考示例。由于要求算法系统具有可解释性所带来的成本数据尚不可知，故需要联合技术专家对解释义务衡量标准予以动态评估以适应不断变化的技术环境。〔87〕从长远来看，精准和动态化的治理还需要以全面准确的算法决策信息集成为治理基础，以科学严谨的信息挖掘为治理前提，以相宜有效的治理需求为治理指针，对算法治理能力有机再造。〔88〕因此，已经有理论

〔85〕 Ibid., p. 843.

〔86〕 USACM, “Principles for Algorithmic Transparency and Accountability”, January 12, 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf., last visited October 10, 2019.

〔87〕 See Finale Doshi-Velez et al, supra note 72, p. 10.

〔88〕 参见李大宇：“精准治理：中国场景下的政府治理范式转换”，《公共管理学报》2017 年第 1 期，第 1—13 页。

和实践将算法影响评估制度作为配合算法解释机制的重要设计,〔89〕通过明确评估要素清单奠定动态化、精细化的治理基础。从更为长远的视角来看,安东尼·凯西等人还提出借助大数据和算法技术,由立法者运用洞察力和前瞻性决策能力,将统一化的法律制度转变为个性化的法律(personalized law),实现法律制定和法律实施的语境化和精确化。〔90〕但无论采用何种方式,都应当意识到不同应用场景下建立多元化、层次化算法解释衡量标准的必要性,其不仅有助于防止解释权流于形式并产生逆向影响,〔91〕还可以为技术创新奠定积极的制度基础。

(四)为算法解释权的有效实施奠定协同治理机制

算法治理是涉及多元主体间复杂互动的议题。在这一现实挑战下,应当意识到算法解释权的有效实施需要以适宜的配套制度作为基础协同推进。在聚焦于算法解释权和算法解释义务构建的同时,立法者还应在权利建构思路之外探索一种兼具内部和外部视角、法律和技术有机结合的协同机制以获得制度实效最大化。通过前文对 GDPR 算法条款的系统性分析,结合算法解释技术客观原理,这里提出协同治理机制供我国立法参考和借鉴。

第一,培育控制者和处理者的内部算法治理机制。大数据时代,有效的治理机制需要遵循信息控制者的动机,培育激励相容的内生机制,调动多方参与治理的积极性。〔92〕算法解释是实施算法问责的重要机制。〔93〕让算法主体承担可问责义务不仅需要其承担证明算法决策正当和准确的义务,还意味着其应努力消除负面社会影响和潜在危害。〔94〕AP29 指南建议,企业应当构建有效的内部监督机制,对个人将产生重大影响的算法应当向内部独立的数据保护官提供影响评估的相关信息。〔95〕同时,企业内部技术团队应当配备专业权威人员对系统的准确性负责,确保其信息可被公众获得,并为救济制度随时启动奠定基础。〔96〕从算法偏误产生的机理来看,算法治理也在很大程度上依赖于开发者和控制者的内部质量整体控制机制。

---

〔89〕 Sidley Austin LLP, "NYC Automated Decision-Making Task Force Forum Provides Insight Into Broader Efforts to Regulate Artificial Intelligence", Lexology, https://www.lexology.com/library/detail.aspx? g=465dbb86-112d-451f-9217-a0528046c874, last visited October 18, 2019.

〔90〕 See Anthony Casey, Anthony Niblett, "A Framework for the New Personalization of Law", *The University of Chicago Law Review*, Vol. 86, No. 1, 2019, pp. 333-358.

〔91〕 See Finale Doshi-Velez et al, supra note 72, pp. 1-12.

〔92〕 周汉华:"探索激励相容的个人数据治理之道——中国个人信息保护法的立法方向",《法学研究》2018 年第 2 期,第 12-15 页。

〔93〕 AI Now Institute, "Algorithmic Accountability Policy Toolkit", October 2018, https://ainowinstitute.org/aap-toolkit.pdf, last visited October 20, 2019.

〔94〕 Nicholas Diakopoulos, Sorelle Friedler, "How to Hold Algorithms Accountable", *MIT Technology Review*, November 17, 2016. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/, last visited October 15, 2019.

〔95〕 See Margot Kaminski, supra note 31, pp. 215-217.

〔96〕 See Nicholas Diakopoulos et al, supra note 94.

算法偏误主要由认知偏误和统计偏误导致。该两类偏误存在时,数据个体很难独立有效地予以识别,而设计者通过对总体样本进行统计和观察,衡量系统整体绩效能提高发现偏误的概率。〔97〕因此,数据控制者和处理者是深入算法设计结构和运行架构的首要主体,算法解释权的具体行使有赖于内部监督机制的协同配合。

第二,从算法可解释性走向具备可解释性的算法模型。欧盟立法虽然对算法解释权进行了布局,但对于具备可解释性的算法模型只有原则性规定。但算法技术的运行机理需要立法者提供充分的技术开发激励,推进具备可解释性算法设计的实现。近期发表于自然杂志子刊的研究表明,目前对具有黑箱特性的算法模型进行解释时仍然受限,依据部分特征展开的解释无法保证可信性、还原性。〔98〕众所周知的 ProPublica 对于 COMPAS 软件发布的报告就是一个经典例证。ProPublica 的报告本质上体现为外部主体对 COMPAS 软件决策逻辑做出的解释。该报告指出,COMPAS 软件基于种族并结合主体其他特性对被告人进行累犯预测,可能存在种族歧视的算法偏误。〔99〕但该项研究指出,ProPublica 的解释虽然模仿了原始 COMPAS 模型的计算过程,却并非完全忠于原始模型。ProPublica 仅对原始模型的部分特征与预测结果关系进行了趋势汇总,认为其是一个基于种族的线性模型。〔100〕但实际上,COMPAS 是非线性模型,客观上的确存在不依赖种族进行预测的可能性。因此,ProPublica 的报告并非对 COMPAS 的"真正解释",相反会产生极大误导。〔101〕这一例证形象地展示出算法解释可信度的重要政策意涵。因此,应当通过立法明确,在高风险类型的算法决策领域,只要存在具有相同性能的可解释性模型,不应允许黑箱模型被部署和使用。同时,还应以立法方式为黑箱算法设计者和控制者建立责任追究机制,通过积极的制度设计促进机器学习领域商业模式和技术模式的转变。〔102〕

四、结 语

伴随着算法社会的全面来临,算法在我国商业和公共事业场景中扮演着愈益重要的角色。我国算法治理实践需求日益攀升,探索有效的治理方案已属势在必行。在方案设计过程中,算法可解释性与算法解释权始终是立法者无法绕开的核心议题。本文借助算法解释权这一在算

〔97〕 See Finale Doshi-Velez et al, supra note 72, pp. 10－11.

〔98〕 See Cynthia Rudin, supra note 37, pp. 206－215.

〔99〕 Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, "Machine Bias", *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, last visited September 20, 2019.

〔100〕 See Cynthia Rudin, supra note 37, pp. 207－208.

〔101〕 See Cynthia Rudin, supra note 37, pp. 206－215.

〔102〕 See Cynthia Rudin, supra note 37, p. 210.

法治理领域的核心基础概念，深入剖析了其蕴含的三项规范价值，并以 GDPR 为立法样本，系统阐释了算法解释权的构建机理和得失利弊。以理性算法和算法理性的调和统一为根本治理原则，本文深入探讨了算法解释权构建的本土化路径，提出了算法解释权的构建策略、行使要件、衡量标准以及协同配套机制。算法治理是一项复杂而长期的工程，未来还需要对算法问责制度、算法影响性评估、算法监管组织架构等核心议题加以深入研究，以期为有效的算法治理实践做出积极的理论贡献。

Abstract: With the increasingly urgent need for algorithm governance, the right to explanation of automated decision-making has been put forward, which has become a ground for empowering users and relevant individuals with respect to their autonomy, the exercise of their technological due process rights, and the avoidance of externalization and shifting of costs and harm caused by algorithm operations. The General Data Protection Regulation (GDPR) designed the right to explanation of automated decision-making in a limited and weakened manner, but it created a combined and reinforced version for algorithm explanations by recognizing various rights of both the data subject and the data protection impact assessment mechanism. Nevertheless, the GDPR scheme has several limitations, such as the relevant provisions being made in an incomplete structure, leaving ambiguity and inappropriate limiting of the application scope. When establishing a localized version of the right to an explanation of algorithms, the position and function for algorithm governance should be properly clarified. The major elements and contents of the right should be fully identified, and in accordance with the degree of social embeddedness and specific application fields. An accurate and scenario-based scheme for explanatory measurement should also be introduced, followed by an integrated approach to governance built on collaboration between internal and external actors.

Key Words: Algorithm Governance; Right to Explanation of Automated Decision Making; Algorithmic Impact Assessment

(责任编辑：彭　錞)