中外法学 Peking University Law Journal
Vol. 36, No. 4(2024)pp. 866－883

算法公平的类型构建与制度实现

张 欣*

摘 要 伴随智能社会的兴起,算法公平已跃升为人工智能治理最为核心和紧迫的议题。基于主体和技术生命周期,算法公平可被归类为个体算法公平和群体算法公平;起点算法公平、过程算法公平和结果算法公平。以此类型框架为据审视我国算法公平规范现状,可见其呈现出体系化不足、可操作性欠佳、规范维度失衡、治理工具缺失以及对具备通用目的的人工智能与特定人工智能存在治理盲区等问题。为破解算法公平治理之困,应依循算法公平的技术机理和伦理关切,凝练法律、伦理和科技之间的内在共识,将算法非歧视原则作为算法公平治理底线,实现跨领域共识的有机融合。在此基础上,我国应立足反歧视法理,通过动态构建差异化的受保护特征清单、打造具有一致性和可预测性的算法歧视审查框架,建立合法性与必要性并重的算法影响评估机制,探寻算法公平的融贯理路与法治化实现路径。

关 键 词 算法公平 算法歧视 数字正义 人工智能治理 通用人工智能

引 言

人工智能技术日新月异,人类社会的智能化水平正在以前所未有的速度飞速攀升。作为市场和政府之外调控资源的“第三只手”,算法通过分类、排序、过滤、搜索、推荐、预测、评估等技术组合,直接塑造着人们被对待的方式和潜在机会。〔1〕然而,在人类不可逆转地步入数字化生存的进程中,算法不公的风险日益凸显,成为亟待应对的严峻挑战。算法产生的偏见和歧视常被隐藏在代码之中,具有极强的跨域性和“结构锁定性”,系统性地侵蚀个人权益和社会公

* 对外经济贸易大学法学院教授。本文系北京市教育科学“十四五”规划课题(项目编号:3030—0014)的阶段性研究成果。

〔1〕 David Beer, “The Social Power of Algorithms,” *Information, Communication and Society*, Vol. 20, No. 1, 2017, p. 6.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

平。算法公平和数字正义因而跃升成为人工智能治理最为核心和紧迫的议题。无论是我国最新提出的《全球人工智能治理倡议》,还是凸显全球关切的《布莱切利宣言》以及联合国首个关于人工智能的全球决议,〔2〕均将算法公平置于可信向善、以人为本的人工智能治理版图之首。尽管算法公平治理的紧迫性已不容忽视,我国法学界对于算法公平基础理论的研究和关注却尚显不足。现有研究多聚焦于特定领域的治理实践,侧重于从部门法规制和域外立法比较的视角切入。〔3〕这些研究为算法治理实践奠定了理论基础,但基础理论研究的缺失不仅制约了算法公平规范体系的构建,也造成了法律治理与技术实践的断裂。算法公平在法律意义上所蕴含的规范诉求与技术开发者的理解存在偏差。司法裁判中所强调的个案公平和情境敏感性与通过量化指标实现算法公平的技术思维之间亦存在巨大鸿沟。〔4〕更为严峻的是,法律人对缓解算法不公技术措施的有效性寄予过高期望,反而忽视了算法公平立法的紧迫性,相关规则未能及时跟进,最终酿成了实践中"曲直难明、公道难彰"的窘境。〔5〕与此同时,作为科技伦理之首要原则的算法公平也失去了被法律吸纳、确认和转化的机会,导致法律、技术和伦理各行其是的割裂局面。鉴于算法公平治理在实践中的紧迫性与理论基础的薄弱性,本文拟系统梳理算法公平的多元内涵,深度剖析我国算法公平治理规范的现状与不足。在此基础上,立足于法律、技术、伦理三位一体的融贯思路,厘清算法公平的衡量尺度与判据,探寻一套明晰可据的算法歧视认定判准,以期为在智能化转型的浪潮中捍卫数字公民的平等与尊严探索"法治化可能"。

---

〔2〕《抓住安全、可靠和值得信赖的人工智能系统带来的机遇,促进可持续发展》,载联合国官网,https://www.undocs.org/Home/Mobile?FinalSymbol=A%2F78%2FL.49&Language=E&DeviceType=Desktop&LangRequested=False,最后访问日期:2024年6月18日。

〔3〕聚焦特定领域的代表性研究,参见许光耀:"大数据杀熟行为的反垄断法调整方法",《政治与法律》2024年第4期,第17—29页;胡萧力:"算法决策场景中就业性别歧视判定的挑战及应对",《现代法学》2023年第4期,第59—74页。以部门法规制为视角的代表性研究,参见杨玉晓:"人工智能算法歧视刑法规制路径研究",《法律适用》2023年第2期,第86—94页;潘芳芳:"算法歧视的民事责任形态",《华东政法大学学报》2021年第4期,第55—68页。域外规制比较的代表性研究,参见郑智航、徐昭曦:"大数据时代算法歧视的法律规制与司法审查——以美国法律实践为例",《比较法研究》2019年第4期,第111—122页;丁晓东:"算法与歧视:从美国教育平权案看算法伦理与法律解释",《中外法学》2017年第6期,第1609—1623页。

〔4〕Sandra Wachter, Brent Mittelstadt and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law And AI," *Computer Law & Security Review*, Vol. 41, 2021, p. 28.

〔5〕在北大法宝和中国裁判文书网以"歧视""平等权"为关键词查询,可以发现我国目前的算法歧视案件均局限于价格歧视领域,尚无一起可公开查询到的与公民平等权相关的算法歧视案件。但相关研究显示,我国人工智能训练数据及算法中均存在歧视性因素。例如,由联合国妇女署资助支持、玛娜数据基金会发布的《促进人工智能算法性别平等研究报告(2021)》指出我国在搜索引擎、媒体传播、智能招聘、电商消费等领域出现普遍的算法性别偏见与歧视。参见玛娜数据基金会:《促进人工智能算法性别平等研究报告(2021)》,第9—22页,载微信公号"玛娜数据基金会",2021年9月28日上传。

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

一、算法公平的类型与内涵

历经蒸汽革命、电气革命和数字革命之后，海量数据、超强算力、复杂算法的出现点燃了智能革命的星星之火，算法社会全面到来。科技行业的崛起和算法权力的出现促使一部分掌握计算资源和数字应用的阶层跃迁成为“代码精英”。借助算法和平台，他们获得了调控和分配资源的能力，隐而不彰地重塑社会。〔6〕在这一洪流之下，面对层出不穷的算法不公现象，各国监管者正努力探寻将这一强大的新型权力纳入制度轨道的有效途径。无论是《美国算法问责法案》《欧盟人工智能法》还是我国的《个人信息保护法》均明确载有算法公平专条。〔7〕然而，何为算法公平、如何衡量算法公平、如何在科技、法律以及伦理维度融贯算法公平，却一直是一个悬而未决的棘手难题。为探寻算法公平的法治化实现路径，必须首先厘清算法公平的概念谱系。鉴此，本部分拟对算法公平的类型与内涵进行归纳与梳理，为后文剖析我国的算法公平规范体系奠定理论框架。

（一）个体算法公平与群体算法公平

算法的本质是一种预测和决策。当面向被预测和决策的主体时，公平是指不存在基于个人或群体的内在或后天特征的任何偏见、歧视或不公正。在这一维度之上，算法不公是指算法决策对某一个体或者特定群体存在偏见，由此引发对该个体或者群体的不公正待遇，并使其利益受损的现象。〔8〕以个体作为基准，辛西娅·德沃克（Cynthia Dwork）等人提出了算法公平的“金标准”，即算法公平是指相似的人应得到相似的对待。当给定同一任务时，任何两个相似的个体应被相似地分类。〔9〕个体公平的概念虽占有一席之地，但基于以下四个原因，个体算法公平具有局限性。首先，仅评估个体之间是否受到类似待遇难以确保个体获得全面意义上的公平。其次，个体公平的相似性度量需要人类评估，但人类决策存在噪音和隐性偏见，难以作出公允评判。再次，确定个体是否具有相似性需借助对相关特征的确认。而这一过程须引入主观判断，这可能导致循环论证的出现。最后，道德价值难以被通约化。从技术角度来看，难以确保价值判断均被准确无误地聚合或者替代为相似变量，这就导致在技术层面对个体开

〔6〕 See Jenna Burrell and Marion Fourcade, “The Society of Algorithms,” *Annual Review of Sociology*, Vol. 47, 2021, pp. 215-216.

〔7〕 See S. 3572 Algorithmic Accountability Act of 2022, Section 4 (a) 11 (B); Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Document PE 24 2024 REV 1, Recital (27)；参见《个人信息保护法》第 24 条。

〔8〕 参见古天龙、李龙等：“公平机器学习：概念、分析与设计”，《计算机学报》2022 年第 5 期，第 1019 页。

〔9〕 See Cynthia Dwork et al., “Fairness through Awareness,” in Shafi Goldwasser (ed.), *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York: Association for Computing Machinery, 2012, pp. 214-216.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

展相似性度量的效果不尽人意。〔10〕因此还需引入群体算法公平的概念。

与个体算法公平有所不同，群体算法公平是指在一个决策过程或分类器系统中，种族、性别等身份敏感特征与决策结果在统计上是独立无关的。这意味着，隶属于不同群体的个体在相同条件下应当具有平等获得有利决策结果的机会或概率，决策者不应基于个体所属群体施加差别对待。〔11〕群体算法公平的重点在于关注依据不同特征划定的群体所获得的待遇是否与各群体的成员构成相称。〔12〕例如，要求算法对不同种族的人具有一致的假阳性率就是在群体算法公平意义上提出的要求。美国的再犯风险评估算法 COMPAS 就曾因其对非洲裔美国人具有更高的假阳性概率而广受诟病。〔13〕然而，群体算法公平同样难以回避理论和现实的双重挑战。首先，现有的群体算法公平度量标准未臻完善，可能会导致有失公允的反直觉结果，出现了满足算法群体公平标准但仍然对相关个体产生“明显不公平”的情形。〔14〕例如，在申请贷款的应用场景中，虽然设计层面满足了群体算法公平的衡量标准，却可能导致不同群体中信誉度类似的个体在贷款获批概率上呈现相异结果。这意味着在倾向群组概率公平的同时，可能造成个体公平损耗。更为棘手的是，算法部署实践中，群体算法公平所要求的约束条件可能互不相容，形成一个棘手的伦理技术难题。〔15〕一方面，群体公平诉求多元，不仅针对不同受保护群体在准确率、误判率、获益比例等指标提出公平对待的要求，还涉及机会公平、结果公平等不同维度的考量；另一方面，现实世界中不同群体的数据分布和历史表现往往存在显著差异，导致约束条件相互冲突，难以同时满足。因此，如何在群体公平约束条件之间进行权衡取舍是一个涉及算法设计选择、商业伦理考量以及法律政策博弈的复杂优化问题。

（二）起点算法公平、过程算法公平和结果算法公平

从技术生命周期来看，算法公平议题贯穿于算法设计、开发、部署、维护直至处置全过程。相应地，算法公平也呈现出阶段性特点。依据算法生命周期，可从起点、过程和结果三个维度探讨算法公平的要义所在。

第一，起点算法公平。该公平类型聚焦算法的设计和开发阶段，强调数据集的选用和处理应保持公平无偏。鉴于产业链下游对该阶段的高度依赖性，起点公平在算法公平治理中处于源头地位。实践中，训练数据集的选择往往受限于可获得性和可访问性，导致某些代表特定群

〔10〕 See Will Fleisher, “What's Fair about Individual Fairness?” in Marion Fourcade, Benjamin Kuipers, Seth Lazar and Deirdre Mulligan (eds.), *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, New York: Association for Computing Machinery, 2021, p. 487.

〔11〕 See Solon Barocas, Moritz Hardt and Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, Cambridge: The MIT Press, 2023, pp. 54-55.

〔12〕 参见张恩典：“数字接触追踪技术的实践类型、社会风险及法律规制”，《法学论坛》2022 年第 3 期，第 102—103 页。

〔13〕 See Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, “Machine Bias,” in Kirsten Martin (ed.), *Ethics of Data and Analytics: Concept and Cases*, New York: Auerbach Publications, 2022, pp. 254-264.

〔14〕 See Dwork et al., supra note 9, p. 218.

〔15〕 See Fleiher, supra note 10, p. 481.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

体的数据被排除在外，进而产生歧视性决策和系统性偏差。〔16〕算法建模过程还不可避免地对背景知识进行简化处理，导致设计阶段使用替代指标可能遗漏关键信息。以再犯风险预测算法为例，设计者通常选取犯罪记录、社区环境、教育程度等指标作为输入特征。这些指标可能与种族等敏感属性高度相关。由于长期存在的结构性不平等，部分少数族裔群体所在社区的犯罪率往往更高。如果算法仅考虑社区犯罪率这一表征，就会不当忽视再犯形成的深层次社会成因，强化既有偏见，将特定群体推向更为不利的处境。因此，严格把控特征选择和数据处理过程，对于保障起点公平至关重要。

第二，过程算法公平。该公平类型关注算法决策程序的公平性，强调决策所依赖的特征选择应当在道德和伦理层面符合社会接受度，具备正当性。这一视角突破了仅关注结果的思维定式，将算法模型的中间过程纳入公平审查的范畴，为全面把握算法公平提供了新的切入点。过程公平要求输入特征具备个体自愿性、特征可靠性、隐私保护性以及与决策任务的相关性。〔17〕仍以再犯风险预测算法为例，过程公平性要求算法设计者评估用于预测的各类特征在道德和伦理层面符合社会接受度。例如，尽管年龄这一特征可能与再犯概率存在统计相关性，但由于其属于个体的“先赋因素”，即人们出生伊始所具有的人力难以选择和控制的因素，个体对其并无选择余地。因此，将其用于算法设计时需要详细审慎的伦理审查和技术论证。为系统评估算法过程公平性，学者提出了多个衡量指标。例如，可将特征先验公平性、特征准确公平性以及特征差异公平性作为评判标准。〔18〕

第三，结果算法公平。这一公平类型直面资源分配与决策后果，从计算和利益分配两个维度审视算法公平性。计算维度关注不同群体在统计意义上的均等对待，旨在全面审视算法决策的实际效果，揭示隐藏在中立表象之下的潜在偏倚。机会均等、赔率均等指标均聚焦于此。例如，机会均等要求应得机会者不因群体属性差异而在资格判定上受到不平等对待。赔率均等则要求假阳性率和假阴性率在相同的受保护属性的类别中是相等的。〔19〕然而，由于此类评价聚焦表征差异而忽视现实影响，仍难以确保实质公平。因此，结果公平还纳入了对利益分配的关注，对算法决策引致的利益获取和负担分配开展全面评估。以智能医疗疾病预测算法为例，如果仅关注疾病预测在不同群体中的准确率和召回率是否均等，可能会掩盖其对病人实际健康影响的巨大差异。对于致命疾病而言，漏诊（假阴性）可能意味着患者错失早期治疗的

〔16〕 See Reva Schwartz et al., “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,” March 2022, p. 15, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf, last visited on 18 June 2024.

〔17〕 Nina Grgić-Hlača et al., “Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018, pp. 52-53.

〔18〕 Ibid., p. 53.

〔19〕 See Mike Teodorescu and Christos Makridis, “Fairness in Machine Learning: Regulation or Standards?” *Brookings*, 15 February 2024, https://www.brookings.edu/articles/fairness-in-machine-learning-regulation-or-standards/, last visited on 18 June 2024.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

机会，而误诊（假阳性）则可能导致患者承受不必要的医疗干预和心理负担。因此，对不同群体的假阴性率与假阳性率应当进行差别化权衡。对于高致死率但早期治疗效果显著的疾病，降低假阴性率更为紧迫；而对于良性病变，控制假阳性率则显得尤为重要。由此可见，聚焦结果层面的算法公平审查不应拘泥于单纯的统计对等，而应立足具体情境，兼顾对不同群体利益和负担的整体考量，充分考虑现实后果对利益主体的差别化影响。〔20〕

由上可见，算法公平是动态而非静态的，是多维而非单向的，是均衡且易变的。每种公平类型均有其适用的场景与边界，多种公平类型之间还可能存在冲突与张力。〔21〕因此，算法公平治理具有复杂性和艰巨性。技术界所宣称的各类算法公平技术可能仅揭示了有限面向，其背后往往隐藏着模糊甚至矛盾的价值预设。算法公平的实现有赖于夯实法律之治。以清晰明确的法律规范为基石，以个案化的司法论证为利器，有助于在海量复杂的具体情境中抽丝剥茧，厘清算法公平的真谛。

二、算法公平治理的中国实践：现状与特征

算法技术的飞速发展既蕴含着促进社会发展的巨大潜力，也对既有的制度框架提出了新的挑战。面对日益攀升的算法公平治理需求，我国近期发布的《全球人工智能治理倡议》多次提到“公平”二字，分别就“提升人工智能技术的公平性”“发展人工智能应符合公平、正义等全人类共同价值”以及人工智能治理的公平性和非歧视性原则作出了积极阐释。在此背景下，系统剖析我国算法公平治理规范的发展脉络与现实图景具有重要的理论价值和实践意义。本部分拟结合前文提出的类型框架，系统梳理并分析我国与算法公平相关的规范与标准，探究现行治理方案的局限所在，为寻求完善之道提供参考依据。

（一）算法公平规范的碎片化态势凸显

总体而言，我国现行算法公平规范体系呈现出基础法律供给不足、部门规章场景受限、技术标准覆盖面窄的碎片化局限。以《个人信息保护法》为代表的基础性法律虽对算法公平有所关注，但仅局限于价格歧视场景，对算法公平的整体规制尚不充分。与之相对，相关次级立法也主要围绕“大数据杀熟”问题构建规制框架，未能从根本上关照算法公平之丰富内涵。以《国务院反垄断委员会关于平台经济领域的反垄断指南》第17条和《互联网信息服务算法推荐管理规定》第21条为例，两者分别从市场结构和算法推荐服务的视角对“大数据杀熟”进行规制，〔22〕在特定领域内形成了针对性的价格歧视治理机制。然而，这些规范均局限于个性化定

〔20〕 See Hilde Weerts, Lambèr Royakkers and Mykola Pechenizkiy, “Does the End Justify the Means? On the Moral Justification of Fairness-Aware Machine Learning,” 2023, p. 3, https://arxiv.org/pdf/2202.08536.pdf, last visited on 18 June 2024.

〔21〕 See Lily Morse et al., “Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms,” *Journal of Business Ethics*, Vol. 181, No. 4, 2022, p. 1084.

〔22〕 参见董彪：“规范自动化决策遏制‘大数据杀熟’——以个人信息保护法第二十四条为视角”，载《民主与法制时报》2021年11月4日，第6版。

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/

价场景，对于更为重要的、涉及个人平等权以及人格尊严的性别、年龄、健康等敏感属性的算法歧视问题，均付之阙如。〔23〕由此可见，我国算法公平规范体系虽已初具雏形，但仍存在基础制度缺位、治理框架碎片化的问题，难以从根本上回应算法技术带来的治理挑战。对此，亟需从基础制度出发，通过推进人工智能基础立法，系统、全面地构建算法公平治理机制。2023 年以来，《人工智能法》已连续两年被列入国务院年度立法工作计划，并被列为预备提请全国人大常委会审议的立法项目。〔24〕这意味着，作为人工智能治理的基础性、综合性法律，《人工智能法》已被正式提上立法议程。对此，立法者应高度重视算法技术对个人权利的深远影响，在兼顾技术创新、产业促进等立法目的时，将对个人基本权利和人格尊严的保护置于首要位置，明确算法公平原则并为之构建清晰的制度实施构架。具体而言，应面向技术生命周期，围绕算法研发、部署和应用，在契合人工智能产业链堆栈架构的基础上，分层次、系统性地构建算法公平治理机制。以《欧盟人工智能法》为鉴，该法以《欧盟基本权利宪章》中公民基本权利保护为基点，在人工智能基础治理框架中嵌入了贯穿数据、算法、模型、应用全周期的公平治理方案。该法对人工智能系统可能产生的歧视性影响设置了风险评估和预防机制，还创设了外部监督机制和救济途径，以防范人工智能的歧视性影响。〔25〕值得注意的是，该法在序言中明确定位其与既有反歧视法律的补充关系，为两套规范的有机衔接提供了法理基础。〔26〕鉴此，我国的《人工智能法》亦应在开拓算法公平综合治理机制的基础上，在概念界定、审查机制、责任分配、救济方式等方面作出清晰规定，并与现有反歧视法律框架实现有效对接，形成规范合力。

（二）算法公平规范的可操作性和指引性不足

我国现行算法公平规范在可操作性和指引性方面存在明显不足。虽然以《全球人工智能治理倡议》《新一代人工智能治理原则》等为代表的顶层文件提出了面向个体和群体的算法公平原则，但缺乏有效的制度支撑，尚未形成层次分明、可资操作的制度方案。立法层面的欠缺导致司法实践难以发挥应有效用。公众难以依据现有法律对不公的算法系统提起反歧视之诉。诉权基础缺失的背后，折射出那些内含歧视与不公风险的算法系统披上技术理性的外衣，持续运行于现实场景而不受约束的治理困境。以外卖骑手为例，面对屡遭诟病的调度决策类

〔23〕《个人信息保护法（草案）》第 29 条曾明确将可能导致个人受到歧视或者人身、财产安全受到严重危害的个人信息，包括种族、民族、宗教信仰、个人生物特征、医疗健康、金融账户、个人行踪等信息归入敏感个人信息并提出更为严格的信息处理要求，但正式颁布的《个人信息保护法》第 28 条在界定敏感个人信息时删除了与歧视治理相关的规定。

〔24〕《国务院办公厅关于印发〈国务院 2024 年度立法工作计划〉的通知》，载中国政府网，https://www.gov.cn/zhengce/zhengceku/202405/content_6950094.htm，最后访问日期：2024 年 6 月 18 日。

〔25〕《欧盟人工智能法》序言（80）指出，对人工智能高风险系统的界定是在充分考量人工智能系统对《欧盟基本权利宪章》所保护的基本权利造成的不利影响程度的基础上设立的。这些权利包括人的尊严以及不受歧视权、残疾人权利以及性别平等。See Artificial Intelligence Act, Document PE 24 2024 REV 1, Recital (80).

〔26〕《欧盟人工智能法》序言（45）指出，与数据保护、消费者保护、基本权利、就业和工人保护以及产品和安全相关的法律而言，该法是已有法律的补充，其出台并不影响现有法律的适用。See Artificial Intelligence Act, Document PE 24 2024 REV 1, Recital (45).

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

算法，骑手群体仅能依赖社交媒体持续发酵的舆论压力和外卖平台组织的“恳谈会”等非制度性渠道，而难以诉诸法律救济。〔27〕虽然《关于落实网络餐饮平台责任 切实维护外卖送餐员权益的指导意见》明确规定平台应以“算法取中”方式合理构建调度决策类算法，适当放宽配送时限，但由于该规定效力位阶较低、内容宽泛抽象，缺乏可操作性和明确指引性，〔28〕导致司法实践中尚无骑手起诉调度决策类算法的成功案例。面对算法操控的不利处境和低微的议价能力，骑手群体往往只能寄望于平台企业的自律，却无法从源头上通过法律手段改善工作境遇。

相比之下，意大利的骑手群体则通过算法歧视之诉，成功获得法律救济。2019 年，意大利博洛尼亚劳工总联合会运输业劳动工会（Filt Cgil Bologna）等工会组织将意大利户户送有限责任公司（Deliveroo Italia S. R. L.）诉至法院，指控其骑手派单算法具有歧视性。法院经审理认为，被告算法系统对骑手进行绩效评估时，未充分考虑其基于参与罢工、患病、照料子女等合法正当理由推迟取消订单的情形，剥夺了其选择工作条件的权利，构成算法歧视。该案从制度构建层面可以提供四项重要启示：其一，面对算法时代反歧视保护的新需求，传统反歧视法律体系的数字化革新至关重要。法院在该案中援引了意大利此前修订的保护工人自由和尊严的相关法令，认为在平台经济语境下，反歧视规定的适用不应受雇佣关系形式的限制，从而为将反歧视保护框架扩展至从属性较弱的零工经济劳动者，创设了法理基础。〔29〕其二，集体诉讼制度是捍卫算法公平的重要制度支撑。依据意大利 2003 年第 216 号法令，法院采用双重标准审查了工会组织提起集体反歧视之诉的适格性。〔30〕此举意味着，代表性组织获得了对抗集体算法不公侵害的独立诉权，补强了个体诉讼能力的不足。其三，举证责任配置规则的完善是破解算法黑箱困境的关键。法院明确了涉及算法歧视诉讼的举证责任分配原则，即原告提供初步的事实要素，即使是统计数据，如从中可以推测存在歧视行为时，被告即负有证明歧视不存在的责任。法院还进一步指出，平台负有证明其算法运行机制合法合理的责任，而其未能提供相关证据，强化了法院对算法存在歧视性的推定。〔31〕其四，非财产性损害赔偿制度的建立可作为对算法不公施加惩戒，以威慑效应倒逼科技企业重视算法公平的制度杠杆。法院考量了算法歧视行为的持续时间、影响范围等因素，鉴于案件审理时相关算法已被停止使用，判令被告赔偿 5 万欧元，为受害者提供必要救济。〔32〕反观我国现状，个体维权诉求缺乏可资依附的规则基础，算法公平之诉如同“无源之水、无本之木”。赋权劳工群体对算法不公提起司法挑战，是助其跳出“系统化困境”的关键一环。因此，应以算法公平为旨构建具有可操作、可诉

〔27〕 例如，美团制定了《美团（全网）骑手恳谈会实施办法（试行）》。参见《2023 年·美团骑手权益保障社会责任报告》，载美团官网，https://www.meituan.com/csr/people/couriers-development，最后访问日期：2024 年 6 月 18 日。

〔28〕 参见谢增毅：“我国平台用工规制路径的反思与改进”，《中外法学》2024 年第 2 期，第 388—391 页。

〔29〕 See Ilaria Purificato, “Behind the Scenes of Deliveroo's Algorithm: The Discriminatory Effect of Frank's Blindness,” *Italian Labour Law e-Journal*, Vol. 14, No. 1, 2021, pp. 172-174.

〔30〕 Ibid., p. 179.

〔31〕 See Purificato, supra note 29, pp. 187-190.

〔32〕 See Purificato, supra note 29, p. 185.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

性的规则体系，为公众对抗算法不公提供精细化、科学化、精准化的制度供给。

（三）算法公平规范维度失衡

人工智能时代的到来，标志着算法正以一种全新的姿态融入经济社会运行的全过程。作为驱动智能的核心引擎，算法全面发挥着汇聚资源、优化配置、重塑生产和社会的关键作用，由此演化成为智能时代的新型基础设施。算法贯穿于人工智能产业链条的始终，作为“新型中介”，成为打通数据要素与传统生产要素的关键纽带，成为平台型组织架构权力的决定性变量。〔33〕与此同时，算法正沿着平台经济这一突破口加速外溢，全方位向公共基础设施渗透，不断消解传统的社会运行中介，重构人类社会的物质交往和社会交往逻辑，由此上升为支撑整个社会系统运转的基础设施。〔34〕然而，面对算法带来的深远影响，反观我国算法公平规范体系不难发现，我国的算法公平治理存在视野局限。目前，我国虽已有一部部门规章、三部规范性文件、两项技术标准涉及群体算法公平，但其所覆盖的群体面向十分有限。〔35〕即便多部规范性文件明确提出“社会公平”和“社会群体间公平”的概念，但落实到制度层面，仅就新就业形态劳动者的算法公平和民族算法公平问题作出了初步规定。《生成式人工智能服务安全基本要求》虽在附录部分罗列了八项歧视性内容，但对于歧视性内容的具体认定标准、构成要件、衡量判据等核心要素均付之阙如。现有规范在维度上的局限性与算法基础设施属性所要求的治理体系化、制度集成化方向尚有差距，尤其对于算法引发的群体性不公风险关注不足，远不能满足智能社会中保障算法公平的现实需求。事实上，随着算法技术的广泛应用和公平诉求的不断涌现，群体算法公平问题已成为一个不容忽视的现实议题。例如，平台企业可能因住房广告投放算法的歧视行为而面临诉讼和处罚；〔36〕残疾人用户则可能就网约车免费等待时长计价算法的歧视性对待寻求救济；〔37〕女性群体也需要通过诉讼途径来改变金融借贷信用评估

〔33〕 参见胡凌：“论赛博空间的架构及其法律意蕴”，《东方法学》2018 年第 3 期，第 95—96 页。

〔34〕 参见孙萍、王从健、梁慧博：“平台基础设施化的原因、特征与表现”，《青年记者》2024 年第 5 期，第 5—8 页。

〔35〕 分别为《生成式人工智能服务管理暂行办法》第 4 条、《关于维护新就业形态劳动者劳动保障权益的指导意见》（十）、《关于落实网络餐饮平台责任 切实维护外卖送餐员权益的指导意见》、《关于加强互联网信息服务算法综合治理的指导意见》（八）、《信息安全技术 机器学习算法安全评估规范（GB/T 42888—2023）》第 4 条第 1 款和《信息安全技术 个人信息安全规范（GB/T 35273—2020）》第 7 条第 4 款。

〔36〕 See U. S. Department of Justice, “Justice Department and Meta Platforms Inc. Reach Key Agreement as They Implement Groundbreaking Resolution to Address Discriminatory Delivery of Housing Advertisements,” 9 January 2023, https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking, last visited on 18 June 2024.

〔37〕 See U. S. Department of Justice, “Uber Commits to Changes and Pays Millions to Resolve Justice Department Lawsuit for Overcharging People with Disabilities,” 18 July 2022, https://www.justice.gov/opa/pr/uber-commits-changes-and-pays-millions-resolve-justice-department-lawsuit-overcharging-people, last visited on 18 June 2024.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

算法中长期内嵌的性别歧视问题。〔38〕这些鲜活的案例无不凸显群体算法公平治理的现实紧迫性。有鉴于此，我国亟待完善现有的群体算法公平规范体系，以确保受到算法决策影响的弱势群体能够通过正式的制度渠道获得公正对待。

（四）算法公平治理工具体系缺失

算法公平治理具有复杂性，其牵涉数据处理、模型设计、系统应用等技术生命周期的诸多环节，需要一系列治理工具予以有力支撑。细察我国现有规范体系，在公平治理工具的构建上仍存诸多不足。例如，《个人信息保护法》作为基础性法律，其第 24 条既强调自动化决策的透明度，又强调决策结果的公平、公正，彰显出兼顾程序公平与实质公平的立法意图，是重要的上位法依据。然而，相关次级立法在算法公平治理工具的具体设计上仍显单薄，尚未充分回应整个算法生命周期的治理需求，致使上位法的精神理念难以有效转化为可资操作的执行方案。首先，在起点算法公平层面，现有技术标准重点围绕训练数据集的样本数量、样本规模、样本多样性以及数据标注等内容，旨在从源头防范数据偏差，但在外部问责工具的制度供给上凸显不足，有必要在现有规范框架下构建数据公平审查机制形成多元协同的制衡格局。〔39〕其次，在过程算法公平方面，《人脸识别技术应用安全管理规定（试行）（征求意见稿）》虽提出个人信息保护影响评估制度，用以衡量人脸识别算法对个人权益的影响，但在赋权用户参与、保障算法决策可解释等方面仍有较大提升空间，尤其在特征选择等核心环节更需引入伦理审查机制，以推进算法“从工具理性到沟通理性的范式转变”。〔40〕对此，我国有必要在现有规范基础上进一步明确算法公平审查制度，为利益相关方参与算法公平治理提供制度化参与渠道。最后，在结果算法公平方面，我国现有规范多侧重于原则性地强调算法结果的公平、公正，禁止人为操纵算法的自然计算结果，但对于如何评估算法结果的公平性，尚无可供遵循的指标体系。〔41〕换言之，即使出现了明显的算法歧视后果，除价格歧视场景外，依据现有规范仍难以对算法服务提供者有效问责。对此，亟需结合人工智能技术特点和应用场景，探索构建算法公平评估工具，引导算法开发应用主体在设计和部署阶段自觉校正纠偏。

（五）具备通用目的的人工智能和特定人工智能存在公平治理盲区

人工智能技术日新月异。具备通用目的的人工智能和特定人工智能系统逐渐成为算法公

〔38〕 See “Class Action Accuses Apple, Goldman Sachs of Discriminating Against Married Women Who Apply for Apple Card,” *Mpelembe*, 19 April 2023, https://mpelembe.net/index.php/class-action-accuses-apple-goldman-sachs-of-discriminating-against-married-women-who-apply-for-apple-card/, last visited on 18 June 2024.

〔39〕 参见《人工智能 深度学习算法评估规范（AIOSS—01—2018）》第 3 条第 5 款、《生成式人工智能服务安全基本要求》第 5 条第 1 项、《信息安全技术 机器学习算法安全评估规范（GB/T 42888—2023）》第 5 条第 1 款第 2 项第 a 目以及附录 A 第 2 条第 3 款第 2 项。

〔40〕 See Ada Lovelace Institute, “Going Public: Towards a Public Participatory Approach for AI,” 12 December 2023, https://www.adalovelaceinstitute.org/report/going-public-participation-ai/, last visited on 18 June 2024.

〔41〕 参见《个人信息保护法》第 24 条、《信息安全技术 机器学习算法安全评估规范（GB/T 42888—2023）》第 6 条第 2 款以及附录 A 第 2 条第 3 款第 2 项。

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/

平治理的核心焦点。然而,我国现有的算法公平规范体系对于这两类人工智能所蕴含的歧视风险尚未给予充分关注和有效回应,形成了治理盲区。就前者而言,其具有高技术复杂性和产业链影响力。一方面,其通常采用自我监督、无监督学习等前沿范式,在海量数据中训练百亿甚至千亿级参数,技术黑箱性远胜从前。技术开发者对其决策的可控性和可预期性骤降,一旦滋生不公风险,技术矫治的难度更大、效果甚微。另一方面,具备通用目的的人工智能以其卓越的通用性和多任务处理能力可广泛渗透于整个产业链。倘若其作为基座模型对个体或群体权益存在潜在威胁,不公风险将迅速外溢,席卷整个产业生态,酿成牵一发而动全身的系统性公平危机。〔42〕更为棘手的是,当具备通用目的的人工智能技术内嵌于人工智能体系统时,传统的算法公平治理框架恐更难以为继。人工智能体决策依赖于大模型映射的内置知识库和动态感知数据,其自身往往存在难以察觉却根深蒂固的数据偏差,这些偏差可能随着智能体的持续学习而被放大强化。与此同时,人工智能体的角色配置和交互设计过程也可能引入全新的偏见维度,开发者有意无意间注入的刻板印象将通过人格化的方式渗透到智能体行为之中。〔43〕这些偏见交织于智能体的认知架构和交互逻辑之中,且呈现出个性化、动态性和情境依赖性特质,极大地提升了矫治难度。传统的面向特定任务和特定场景的算法公平治理方案恐不足以应对这些前沿尖端科技带来的治理挑战,亟需构建更为灵活、实时、精细的治理框架。

在特定人工智能领域,以情感识别和生物识别技术为代表的系统极易引发歧视性后果。这类系统试图从生物特征推断主体的情绪和意图,但可靠性不足、特异性匮乏、泛化性受限。一旦将其草率部署于就业、教育等高风险场景,极易加剧固有偏见。〔44〕对此,现有规范体系尚未建立针对性的治理框架。未来应着重对这两类人工智能的差异化特征和风险表征给予规制上的精准回应,在总体治理框架下强化技术设计者和服务提供者的算法公平义务,提升治理的针对性。

三、算法公平的融贯理路与实现路径

伴随着算法对人类社会的全方位渗透和深度重塑,其俨然成为塑造现实世界的“隐形建筑师”,深刻影响着个人权利、群体福祉和社会公平的实现。本质上,算法是一个涉及多步骤的决策生成过程,每一个细微的技术抉择都可能影响其公平性。对此,单纯寄望于科技企业的自我规制或对空泛伦理原则的表面追随,难以为算法公平提供充分的制度保障。因此,亟需法律适时介入,发挥规范、引导和矫正功能。然而,传统的反歧视框架旨在防范人类主观意图驱动的显性歧视,对于更为抽象、微妙、隐形、动态的算法歧视往往力不从心。〔45〕因此,从法律视角

〔42〕 See Artificial Intelligence Act, Document PE 24 2024 REV 1, Recital (97) and (110).

〔43〕 参见张欣:“论人工智能体的模块化治理”,《东方法学》2024 年第 2 期,第 132 页。

〔44〕 See Artificial Intelligence Act, Document PE 24 2024 REV 1, Recital (44).

〔45〕 See Wachter et al., supra note 4, p. 10.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

出发，因循算法公平的技术机理和伦理关切，尝试构建一个清晰、一致、可预期的制度框架，具有至关重要的意义。

（一）凝练法律、伦理和科技之间融贯的内在共识

实现算法公平的首要任务是在规范层面和体系层面实现法律、伦理和科技三者的“融贯”。所谓“融贯”，即是从这三个领域中提炼出最为关键的基本共识，使三者形成相互支撑、彼此证立的网络化结构关系。“融贯”的目的旨在超越法律、伦理、科技相互分立的思维范式，在规则层面实现三者的耦合互动与协同共治。由此观之，算法的非歧视原则是构成其内在共识的关键一环，是实现法律、技术和伦理融贯的逻辑起点和关键支点。

作为平等理念的延伸与扩展，算法公平与算法非歧视原则具有内在的价值契合性。平等权作为一项基本权利，是彰显人之为人的价值所在，而歧视恰是对这一价值的根本否定。〔46〕歧视的本质是在某些社会层面对个体或群体施加并无正当理由的、道德上令人反感的、更为不利的差别对待。〔47〕无论是以传统形式还是以数字化形式出现，歧视行为都直接侵犯了人的平等权，违背了人的尊严并侵蚀人的主体性。由此观之，算法公平与算法非歧视原则同处平等权保护这一价值的统辖和指引之下。算法公平治理既是对反歧视传统的延续和拓展，更蕴含着智能时代实现公平的新诉求。算法公平与算法非歧视虽各有侧重，但二者殊途同归，在根本上都指向以平等理念约束算法的开发和应用，防止和纠正自动化决策嵌入和造成非正义差别。因此，算法公平治理不能简单另起炉灶，而应将现有反歧视法律体系的革新纳入考量，在公平理念的指引下，以协调法律规范意义上的平等权保护和创建治理意义上的算法不公防范机制为路径，实现二者的有机融合。

前文对我国现有算法公平规范的分析表明，当前阶段的算法公平治理亟需进行体系化重构，围绕算法公平的多元类型配备与之契合的专门制度。与此同时，还应遵循底线逻辑，从法律层面清晰勾勒出算法歧视的认定框架，将其视为算法非歧视原则落地的制度抓手，以此撬动法律规范、科技伦理和技术设计三个领域的深度融合。值得注意的是，我国算法公平规范的碎片化、原则化趋势并非孤立存在，而是与我国的反歧视法律体系的整体发展状况相互映照。究其根源在于，我国反歧视法尚处发展阶段，与歧视相关的法律定义多借鉴国际条约，诉诸道德直觉表达，缺乏清晰的内涵与外延，〔48〕导致被歧视者在实践中获得法律救济的成本高昂、障碍重重。〔49〕现实制度的局限向数字空间投射，使得算法公平规范体系陷入无力匮乏的窘境。因此，算法公平虽是人工智能时代的新兴治理议题，其应对之道还需溯源制度之本，仰赖于反

〔46〕 See Hilde Weerts et al., “Algorithmic Unfairness Through the Lens of EU Non-Discrimination Law: Or Why the Law is Not a Decision Tree,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, 2023, p. 806.

〔47〕 Ibid.

〔48〕 李志颖：“论歧视的法律定义——基于社会行为视角的分析”，《法制与社会发展》2023年第1期，第128页。

〔49〕 参见阎天：“反就业歧视法的一般理论——中美两国的建构与反思”，《环球法律评论》2014年第6期，第60页。

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/

歧视法律规范的丰盈。

但算法歧视并非传统歧视的简单数字化。算法歧视无需明确歧视意图，却可造成不容忽视的负面影响；同样的算法模型，应用于不同场景可导致迥异的歧视风险；算法歧视成因错综复杂，需依循其技术机理和产业特点探寻精细化归责机制。传统反歧视法领域化、分割化的局限，难以应对算法歧视的跨域性影响。因此，围绕算法非歧视原则构建的治理框架需在反歧视规范基础上展开系统性、针对性、全局性的革新。而动态构建差异化的受保护特征清单以及构建具有一致性和可预测性的算法歧视审查框架即是两个至关重要的制度安排。

(二)动态构建差异化的受保护特征清单

区分是人类理解世界的前提，是人类社会运行的基本逻辑之一。区分是形成群体认同、简化信息沟通的必要机制，也是合理配置稀缺资源的重要依据。因此，合理的区别对待具有正当性，有助于实现认知效率、社会秩序和分配正义。〔50〕由此可见，区分本身并非分配不公的根源，关键在于如何界定恰当的区分标准。正如彼得·韦斯顿(Peter Westen)所言，如果不指明是何种因素使得人或者待遇相似，就无法言说何为平等。〔51〕换言之，实现算法公平的首要问题是厘清何种因素构成了不正当的差别对待。

纵观我国现行规范，《生成式人工智能服务管理暂行办法》《生成式人工智能服务安全基本要求》以及《信息安全技术 个人信息安全规范(GB/T 35273—2020)》均采取“一刀切”方式，简单罗列民族、信仰、国别、地域、性别、年龄、职业、健康、社会阶层等特征，将其笼统纳入受保护特征清单之中。这种线性移植传统反歧视法的做法，既无法实现人工智能场景下对个体和群体平等权的精准保护，也无益于人工智能产业的促进与发展。

首先，受保护特征的确立缘于立法者的价值判断，反映出特定社会语境下对某些主体合法保护需求的认知。〔52〕受保护特征犹如反歧视体系的“定位器”，在特定历史和社会语境下，成为防止个体或者群体因污名、偏见、刻板印象等根深蒂固的原因使其免受基于特定属性歧视的重要连接点。〔53〕因此，简单将传统反歧视领域的受保护特征移植到人工智能治理领域，无法对个体和群体形成有的放矢的保护。相反，过多地、不恰当地识别人工智能领域的受保护特征，客观上会限定训练数据集的充分使用。而如前文所述，数据集的多样化、包容性恰是算法起点公平的题中之义。是以，一刀切式的笼统立法反而不利于算法决策的公平性和准确性。其次，人工智能部署场景中，受保护特征呈现动态性与差异性。例如，同样是性别和种族，在金融定价和授信场景中将其纳入考量可能带来群体性不公平，但在医疗场景下，性别、年龄、种族等敏感属性对疾病诊断和治疗方案的制定却具有重要价值，将其用于差异化的专业判断具有

〔50〕 See Weerts et al., supra note 46, p. 808.

〔51〕 See Peter Westen, “The Empty Idea of Equality,” *Harvard Law Review*, Vol. 95, No. 3, 1982, pp. 537-596.

〔52〕 See Kate Malleson, “Equality Law and the Protected Characteristics,” *The Modern Law Review*, Vol. 81, No. 4, 2018, pp. 599-600.

〔53〕 Ibid., pp. 599-621.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

正当性;但若将这些敏感属性与医疗资源的分配相关联,又可能加剧对特定群体平等权的损害。〔54〕此外,传统反歧视法未涵盖的特征,在算法决策中依然可能交叉重叠、产生新的歧视风险,而简单搬运既有清单,将使新型歧视风险失于识别和应对。〔55〕

由是观之,"受保护特征清单"的制定是实现算法公平在法律、技术、伦理多维融贯的关键一环。基于算法公平理念制定的"受保护特征清单"应统筹考虑垂直行业和高风险场景的差异化需求,吸收现有算法公平技术标准的通用要求,在"定制化"与"标准化"之间寻求平衡。具体制定时,还需注意以下几点:首先,应全面梳理高风险人工智能部署的核心场景,明确其功能定位、运行逻辑等关键特征。不同领域因业务性质、行业规则、数据来源的差异,对平等权的潜在影响不尽相同。目前,机器学习特征选择和模型训练的公平性评估已具备一系列工具、流程和技术标准,〔56〕其探索与实践对制定高风险领域的受保护特征清单具有借鉴意义。其次,深入分析高风险人工智能应用场景的决策机制和信息流动特点。算法因功能定位和决策依据的不同,对特定受保护特征的敏感程度也会有所差异。这就要求在权衡受保护特征时"对症下药",针对性地识别高风险场景下的主要风险点。最后,受保护特征清单应定期更新,确保其与技术发展的协调性。算法公平治理是一个多目标优化的过程,公平与效率、安全与创新等目标间难免存在交织博弈。因此,制定受保护特征清单还须多维权衡、与时俱进,必要时主动纳入新的特征维度,剔除不再敏感的特征类型,为动态调整预留空间。

(三)探寻具有一致性、可预测性的算法歧视审查框架

与传统歧视不同,算法歧视的发生机制具有复杂性。首先,算法歧视并非以因果关系形成决策和输出,个人难以通过直观感知的方式获得初步证据从而启动救济。如我国闫佳琳案中,劳动者可明确得知被告系因"河南人"这一地域事由对其施加差别对待的歧视情形,在算法歧视领域十分鲜见。〔57〕其次,算法歧视多源于数据或者社会现实中已经存在的结构性偏差。算法设计者的主观故意并非必要构成要件。因而,在我国反歧视诉讼中仅接受以直接证据证

〔54〕 See Tiago Pagano et al., "Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach," *Big Data and Cognitive Computing*, Vol. 7, No. 1, 2023, p. 2.

〔55〕 See Jared Council, "LinkedIn Unveils Tool to Help Combat AI Bias," *The Wall Street Journal*, 25 August 2020, https://www.wsj.com/articles/linkedin-unveils-tool-to-help-combat-ai-bias-11598371200, last visited on 18 June 2024.

〔56〕 See the Institute of Electrical and Electronics Engineers (IEEE), "IEEE Standard Model Process for Addressing Ethical Concerns during System Design," 15 September 2021, https://standards.ieee.org/ieee/7000/6781/; the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), "ISO/IEC TR 24028:2020 Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence," May 2020, https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24028:ed-1:v1:en; ISO and IEC, "ISO/IEC TR 24027:2021Information Technology—Artificial intelligence (AI)—Bias in AI systems and AI aided decision making," November 2021, https://www.iso.org/standard/77607.html, last visited on 18 June 2024.

〔57〕 闫佳琳诉浙江喜来登度假村有限公司平等就业权纠纷案,最高人民法院第185号指导案例(2022年7月4日发布),浙江省杭州市中级人民法院民事判决书,(2020)浙01民终736号。

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

明歧视的证据规则势必掣肘算法歧视救济之诉的启动。仅接受直接证据证明歧视的背后逻辑,乃是认为构成歧视的主观要件应为故意,诉讼主体理应证明对方存在过错。〔58〕但如前所述,人工智能语境下,更新证据类型、改造证据规则,适时转移证成责任,理应成为法院积极采用的证据规则。这一转变的深层逻辑在于,法律不仅应保护个人和群体免受基于故意的算法歧视,更应提供制度性保障审查那些看似中立的算法运行规则,从而重塑个人与算法设计者、算法服务提供者之间的权力义务关系结构,以实施国家保护义务防止算法权力对个体和特定群体不当施加差别对待。〔59〕换言之,法律作为一种社会评价和调节机制,应秉持实质公平理念积极介入,扮演算法公平的守望者。这一目标的实现尚需以下规则作为支撑:

第一,构建衡量受保护特征与争议特征变量相关性的分析框架。实践中,算法设计者常规避使用法律禁止的受保护特征变量,转而使用与受保护属性"密不可分"的代理变量。例如,虽然核定保险费率的算法未直接基于"种族"作出决策,但基于特定国家种族聚居的特点,邮政编码仍可达到种族区分的效果。〔60〕但如何认定"密切相关"并构成不正当的区别对待,需要法官借助广泛的社会背景、专业知识以及融合个案情境综合确定。例如,怀孕与性别具有天然的生理性因果关系,但出生地与种族之间未必存在必然关联。不过,考虑到不平等的成本和负担应在算法设计者、潜在受害者和整个社会间合理分配,〔61〕更为适宜的方式是,一方面允许原告提供初步证据、表明决策结果存在显著的统计学差异作为开启诉讼的前提,〔62〕另一方面通过立法前置性地要求算法设计者或者算法服务提供者就决策设计的合理性、必要性和正当性进行披露,同时应具体之诉,就潜在特征变量与受保护特征的相关性予以说明。考虑到用户群体知识、精力和资源的有限性,证明算法决策损害阈值的初步证据不宜设定过高。例如,原告提供了决策结果对个体或者特定受保护群体产生了统计上显著的不利影响,或者是与不利处境直接相关的证明,均宜认为满足举证要求。2022 年,一名荷兰学生曾对所在大学提出投诉,诉称在线考试期间使用的反作弊软件中嵌入的人脸识别算法,因其深色肤色而失败率更高。该学生提出的不利处境证据就是依据人脸识别算法性能差异的科学研究,将其作为该项算法造成基于种族的间接歧视的初步证据之一。〔63〕

对于算法设计者和算法服务提供者而言,建议通过立法规定其事前披露义务,并结合个案要求其履行对应的证明责任。一来,这一方案更契合科技企业所具有的"算法权力"。鉴于算法权力的潜在恣意性,算法设计者理应承担更高的注意义务,在设计环节履行充分合理的算法公平保障义务,对可能与"受保护特征"构成关联的决策变量施加更为审慎的关注。二来,受限于专业知识的缺乏,用户在初始证据收集和举证时,往往难以全面获取算法模型的内部信息和

〔58〕 阎天,见前注〔49〕,第 70 页。

〔59〕 参见王锡锌:"个人信息国家保护义务及展开",《中国法学》2021 年第 1 期,第 157—165 页。

〔60〕 Alexandra George, "Thwarting Bias in AI Systems," 11 December 2018, https://engineering.cmu.edu/news-events/news/2018/12/11-datta-proxies.html, last visited on 18 June 2024.

〔61〕 See Weerts et al., supra note 46, p. 807.

〔62〕 See Weerts et al., supra note 46, p. 811.

〔63〕 See Weerts et al., supra note 46, p. 811.

北大法律信息网www.chinalawinfo.com/ 北大法宝www.pkulaw.com/

运行机制。而算法设计公平性义务的履行情况，恰恰是用户形成初步证据、启动诉讼程序所需的关键信息。通过立法明确事前披露义务，可以削减信息不对称对用户权益保障的不利影响，为用户获取救济提供必要的制度支持。三来，正如路易斯·卡普洛(Louis Kaplow)所言，传统反歧视诉讼中要求法官先收集证明损害阈值的信息，而后才考虑证明利益阈值信息的规则，忽略了证据的协同效应，违反了最优信息收集的原则。这是因为某些证据可能同时与损害和利益具有相关性。〔64〕因此，将事前披露与事后应诉相结合的方案更契合算法歧视的发生机制以及信息收集的最佳原理。

实践中，一些人工智能领域的立法业已采纳了事前披露义务的思路，对算法公平评估所需的特定信息收集义务作出前置性规定。例如，《伊利诺伊州人工智能视频面试法》规定，仅依靠人工智能分析视频面试来决定申请人是否进入面谈的雇主，必须收集并报告两类人口统计数据：①使用人工智能分析后，获得或未获得面谈机会的申请人的种族和族裔；②被录用申请人的种族和族裔。〔65〕再如，2019 年颁布的《欧盟促进商业用户使用在线中介服务的公平和透明度条例》第 5 条和第 7 条前置性地规定，在线中介服务提供商应以简单易懂的语言列出决定排名的主要参数，以及这些参数相对于其他参数更重要的原因。在线中介服务提供商应说明对其利用在线中介服务向消费者提供商品或服务的任何差别对待行为，以及其在经济、商业或者法律角度的主要考量。〔66〕

第二，构建算法决策影响评估制度。算法设计过程中，特征变量的选择与模型性能息息相关。依据不同的任务目标、数据特点、领域知识和模型复杂度，技术界存在各种旨在优化机器学习性能的特征选择方法。〔67〕因此，特征选择是一个涉及多目标优化的复杂过程。这意味着对算法决策影响的评估应力求平衡平等权保护这一重要法益与技术创新的容错阈值。对于争议变量而言，需要评估其是否具有区分对待的“正当理由”。如果某些特征变量的使用所致歧视影响可能超过实现目的的必需限度，则应认定为不具备区分对待的正当性。在具体裁定时，可重点考虑下列因素。

其一，合法性，即特征选取和数据处理方式的目标应具有合法性。这需要结合算法设计初衷以及具体部署场景具体判断。实践中，一项算法虽可能最终被认定为歧视性决策，但其目标可能依然具有合法性。司法实践中，法官应结合独立专家意见，依据具体语境还原算法决策规则的主观预设和价值判断，进而对争议算法的合法性目的作出判断。

---

〔64〕 See Louis Kaplow, “Balancing Versus Structured Decision Procedures: Antitrust, Title VII Disparate Impact, And Constitutional Law Strict Scrutiny,” *University of Pennsylvania Law Review*, Vol. 167, No. 6, 2018, pp. 1387-1389.

〔65〕 See Artificial Intelligence Video Interview Act 2019, 820 ILCS 42, Article 20.

〔66〕 See Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on Promoting Fairness and Transparency for Business Users of Online Intermediation Services, Article 5 and 7.

〔67〕 参见卢泓宇等：“卷积神经网络特征重要性分析及增强特征选择模型”，《软件学报》2017 年第 11 期，第 2881 页。

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/

其二，必要性，即该方式是否为实现合法目标所必需，且无更优替代方案。实践中，对必要性的判断往往最具挑战。司法实践中，部分法院采取较为宽松的标准，只要被告可证明最低限度的效率提升，即认为满足必要性标准。而另一些法院则持严格解释标准，认为被告必须证明该实践带来的效率收益对具体运营至关重要，缺少其无法继续实现决策目的。多数法院则取中间立场，认为实践产生的效率收益应达到一定的实质门槛，明显高于其潜在歧视影响，方符合“必要性”标准。〔68〕这种“结构化判定程序”不仅使抗辩门槛模糊不清，亦与在算法公平与技术创新之间取得良好平衡的目标相悖，存在内在逻辑缺陷。因此，应从利益衡量的基本逻辑出发，判断歧视影响与效率收益孰轻孰重。诚然，后一方案面临的更大挑战在于，如何构建具有一致性和可预测性的“通约化框架”，以比较算法决策的损害阈值和利益阈值。〔69〕但这一挑战并非不可化解。如前所述，衡量各类算法公平多表现为对特定维度的相似性度量。因此，可依据行业内具有共识性的技术标准，制定算法公平的阈值和基准，用以辅助衡量决策中的合理因素和特征，据此确定基本判定方向和考量因素。〔70〕例如，在大学招生录取算法中，可能需要关注不同专业项目的竞争力差异，应分别衡量各专业内部的录取率是否均衡，而非简单比较男女生的总体录取率。〔71〕此外，还可尝试通过模型测算不同类型歧视行为的社会成本、共性效率收益，形成对算法决策社会净收益的参考依据，并结合受保护特征清单以及专家证人的协助，厘清技术细节，评估具体的设计规则。这些配套制度均可发挥补足作用，为构建一致性、可预测性的评估框架奠定制度基础。

四、结　语

智能社会的来临，标志着人类正在经历一场范式革命。算法作为驱动人工智能系统运行的智能引擎，正以前所未有的广度和深度重塑人类的生活方式、生产方式和治理逻辑。其在展现出巨大潜力的同时，也不可避免地带来了技术风险和伦理挑战，其中最为突出和紧迫的，莫过于算法公平治理问题。算法不公的出现源于技术中性的迷思，根植于社会偏见的投射，体现为数字化的不公正对待，影响着个体权益的保障和社会正义的实现。正如图灵奖得主杨立昆所言，在未来，“每个人与数字世界和知识世界的互动都将由人工智能作为中介”。〔72〕这警醒我们应以更加审慎的态度直面算法公平的挑战。“认真对待算法公平”关乎每一个数字公民的切身利益，更关乎整个社会的良性运转。在算法深刻重塑社会形态和权力结构的数智时代，捍

〔68〕 See Kaplow, supra note 64, pp. 1420-1437.

〔69〕 See Kaplow, supra note 64, pp. 1459-1462.

〔70〕 See IEEE, supra note 56.

〔71〕 See Weerts et al., supra note 46, p. 813.

〔72〕 See Lex Fridman Podcast #416, “Transcript for Yann Lecun: Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI,” https://lexfridman.com/yann-lecun-3-transcript, last visited on 18 June 2024.

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/

卫平等、弥合数字鸿沟、消解结构性不公，是法律人的使命担当，也需要全社会的广泛参与和普遍觉醒。

Abstract: With the rise of artificial intelligence in society, algorithmic fairness has emerged as the most crucial issue in AI governance. Based on the subject and technical life cycle, algorithmic fairness can be classified into individual and group algorithmic fairness, as well as fairness at the starting point, in the process, and in the outcome. Examining China's current norms using this typological framework reveals problems such as lack of systematization, weak operability, imbalanced regulatory dimensions, lack of governance tools, and regulatory blind spots for general and specific AI. To resolve the dilemma, it is necessary to distill the inherent consensus among law, ethics, and technology, and establish the principle of non-discrimination as the foundation of algorithmic fairness governance. Grounded in non-discrimination norms, a dynamic and differentiated list of protected characteristics should be constructed, a consistent and predictable commensurable review framework should be created, and an algorithm impact assessment mechanism that emphasizes both legality and necessity should be established for achieving algorithmic fairness.

Key Words: Algorithmic Fairness; Algorithmic Discrimination; Digital Justice; AI Governance; Artificial General Intelligence

（责任编辑：高　薇）

北大法律信息网www.chinalawinfo.com/　北大法宝www.pkulaw.com/