DOI: 10.11817/j.issn. 1672-3104. 2026. 01. 004

迈向算法正义：算法歧视的社会建构及其治理策略

毛俊响[1]，郭敏[2]

(1. 中南大学法学院，湖南长沙，410083；
2. 中南大学人文学院，湖南长沙，410083)

摘要：算法作为一种社会存在，与社会结构相互作用。算法歧视是社会建构的结果，社会歧视的历史承继、算法运行的修正障碍、价值偏好的隐性渗透、社会生态的利益导向，共同作用于算法歧视。算法歧视并非算法内生技术性问题衍生的新形式不平等，而是历史与现实问题在算法时代的映射。算法歧视从财富、权力、声望三个方面阻碍合理化社会流动，加剧了社会结构失衡。为此，需要建构算法正义来实现对算法歧视的纠正。建构算法正义，需要回应分配正义与关系正义的双重要求，将受保护特征纳入算法决策，尊重群体多元、避免多样性“武器化”，正视基于群体差异的特殊优待并提升少数群体的算法决策话语权。实现算法正义，应采用技术、法律、伦理三元协同治理模式，重点是设计以算法区分为中心的法律规制模式，布局以“科技向善”为核心的人工智能全周期治理机制。

关键词：算法歧视；算法正义；算法区分；算法治理

中图分类号：D994 **文献标识码**：A **文章编号**：1672-3104(2026)01−0032−16

一、问题的提出

当前，由人工智能引领的新一轮科技革命和产业变革方兴未艾，以算法和大数据为驱动的人工智能革命被认为是推动边缘化群体公平获取社会资源的重要机遇[1]。但近年来，现实愈发证明，在为人类社会带来潜在巨大发展红利的同时，人工智能摇身一变成为社会歧视的推动者，很大程度上加剧了社会不平等。联合国秘书长古特雷斯指出，数字技术是一个以男性为主的行业的产物，是歧视和偏见新的来源，“基于不完整数据和设计不当的算法的技术非但没有呈现事实并解决偏见，反而数字化并放大了性别歧视”[2]。荷兰政府在儿童福利认定中采用了欺诈检测的算法系统，后发现该算法导致了成千上万个家庭被错误认定为儿童福利欺诈，其中超半数的受害家庭是具有移民背景的双国籍家庭，内阁 2021 年因此集体辞职[3]。当前，算法歧视①已成为一个亟待解决的全球性问题。算法表面上推动了人人共襄信息社会数字盛举的机会均等化，现实中却保持甚至加深了群体间的发展鸿沟。这不仅会消解国家权力在社会关系治理中的效用，而且筑高了不同群体间阻碍交流融合的壁垒，不利于多元社会认同感的培育。

习近平总书记指出：“要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控。要整合多学科力量，加强人工智能相关法律、伦理、社会问题研究，建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德。”[4]近年来，国内学者对算法歧视

收稿日期：2024−11−04

基金项目：教育部哲学社会科学重大课题攻关项目“习近平总书记关于尊重和保障人权重要论述研究”(22JZD002)；湖南省社会科学成果评审委员会课题“算法决策的公平性困境与法律治理研究”

作者简介：毛俊响，男，湖北黄梅人，法学博士，中南大学法学院教授，国家人权教育与培训基地——中南大学人权研究中心研究员，主要研究方向：国际法，联系邮箱：Tangmao200304@csu.edu.cn；郭敏，女，山东青岛人，中南大学人文学院哲学博士后流动站研究人员，主要研究方向：国际法

问题日益重视，主要关注算法歧视的伦理问题[5−6]、法律规制[7−8]等。也有部分学者对算法歧视的某一领域做针对性研究，例如雇佣中的性别歧视问题[9]。相比之下，外国学者对算法歧视问题的研究起步更早，视角更为多元。综合看来，相关成果可以分为三个方面：第一，通过实证研究揭示算法歧视的普遍性，指出歧视性算法在就业、犯罪预测和社会福利等领域的负面效果[10−11]。第二，算法歧视的技术性治理。此类研究认为歧视性算法源于算法本身设计存在缺陷，应从技术角度在数据集调整、编码设计等方面介入算法运行，以实现决策结果的平等性[12]。第三，跨学科的综合治理。学者们从重视伦理重要性、呼吁伦理与科技的交叉课程推广[13]、扩大弱势群体在科技公司的人员占比[14]等方面提出治理算法的各类措施。

我们认为，技术实践具有社会性，与社会结构之间存在互构作用。引发算法歧视的根本原因不是技术问题，而是社会长期存在的系统性、结构性不平等在算法时代的映射。现有研究虽从多个方面讨论算法歧视的法律治理，但若忽视算法歧视的社会建构性，再多的治理措施也只能是舍本逐末之举。在人工智能引发社会革命性变化的时代，从社会建构视角来深入分析算法歧视与社会结构的交互作用，并在明确算法正义建构标准的基础上探索技术、法律、伦理的三元协调治理思路，具有重要的理论和现实意义。

二、算法歧视的社会建构

马克思主义认为，一切社会存在物都具有社会属性，人是一种特殊的社会存在物。但是，“作为社会存在物的人并不是单线地被社会所规定，人是社会的人，反过来，人也建构并重构着属于自己的社会，社会也是人的社会”[15]。不平等的社会催生不平等的算法，数据和代码及其背后蕴含的秩序与理性，无法脱离特定社会因素的形塑作用。社会现象或社会存在生成于特定的历史和文化背景，在社会互动的作用下被建构和维持。这也意味着，生活在极端阶层差异中的少数群体，甚至因身处其中而无法感知由同样的社会环境塑造出的技术异化问题。巴西学者席尔瓦(Silva)研究指出，种族与社会经济地位的交织使一些群体难以意识到自身所受的技术压迫[16]。巴西《通用数据保护法》(LGPD)的核心原则之一是对个人数据收集的知情同意。一些公司在救济食品分发项目中收集生物识别数据，受救助者在饥饿和对个人信息缺乏了解的情况下同意接受救济，而并未真正考虑后果。这导致了贫困群体不成比例的数据收集和风险的增加。席尔瓦指出，对许多人而言，维持生计的现实压力常常压倒了对算法歧视等抽象问题的关注与抗争[16]。

“技术是人类的技术，是社会中的技术，是人类创造出来的文化形式”[17]，因此，理解算法歧视，除了需要关注参数、编码、数组等算法本身的技术性要素外，还应强调社会因素，即算法在不平等的社会背景下生成，被裹挟和建构成一定程度上的“偏见者”。

(一) 社会歧视的历史承继

算法歧视是指由数据驱动的算法系统在设计、开发、训练和部署过程中体现或加剧的偏见和歧视。有观点认为，歧视是大数据算法的“劣根性”[18]，而披露计算机源代码、增强算法决策程序透明度是降低风险的理想举措[19]。事实上，大数据是关于社会和经济趋势以及人类行为的大量统计信息的积累[20]，其内容本身便呈现出现实社会的既有结构与偏差。在机器学习的过程中，这些源于现实的偏差性数据被直接用于算法模型的训练，并在算法运行中持续被复制和放大。例如，美国信用评分系统并没有特意将种族因素纳入评价指标，但其不公平性长期遭到大众质疑，这是因为该系统所参考和使用的数据是基于固化自奴隶制时期的社会财富分配状况和长期以来不平等的社会公共政策塑造的支付

和借贷记录而形成的。“在许多情况下，对算法决策的批评实际上反映了对数字鸿沟的更广泛担忧，甚至是对不平等社会的普遍谴责。”[21]也就是说，算法歧视并不是随着信息化发展而凭空产生的新问题，实际上是国际社会长久以来存在的各类不平等问题在算法时代的延续。

(二) 算法运行的修正障碍

算法给人的第一印象是高度专业化。一个算法既要包括数字运算、逻辑运算、关系运算，也要包括数据传输指令和赋值运算[22]。高度复杂的算法运行过程使得普通用户不仅在算法设计之初因缺乏专业知识无法参与意见收集，而且在算法被实际运用后也无法知晓其运作规则，因此很难对其修正和更新提出意见，只能被动接受决策结果。

许多学者因此建议技术人员设计更透明或可解释的算法系统，期冀通过算法披露、算法公开打开算法黑箱[23]。诚然，作为实现算法可解释目标的阻碍，透明度的缺乏是导致算法歧视的重要诱因之一，但若完全将实现算法公平的目标寄希望于算法披露，则是不切实际的。一方面，将算法纳入商业秘密保护已成为各国共识。德国、日本等大陆法系国家通过反不正当竞争的方式保护算法[22]。“那些支撑重大技术(运行)的关键底层技术都是专有的，公司往往以封闭方式加以投入，这既是出于竞争的原因，也是为了尽量减少外部操纵。”[24] 2020 年 8 月，《最高人民法院关于审理侵犯商业秘密民事案件适用法律若干问题的规定》将算法列入了技术信息范畴，标志着算法正式成为我国商业秘密保护法的对象。

另一方面，算法的高度复杂性决定了即便科技公司不从商业秘密角度予以抗辩，主动向社会公开其训练数据和程序，我们也难以从中判断算法中是否存在偏见和歧视。部分学者探索从个人对算法的解释请求权角度入手，要求算法公开以实现算法决策的透明化[25]。而现实中，编程的抽象化代码和大众专业知识储备存在巨大鸿沟：“由于大多数大型系统都涉及与数据相关的算法，并根据输入进行进化，因此对它们进行脱离上下文的研究，几乎无法阐明一个人在任何给定时间的搜索结果是什么样子。”[24]还有学者探索建构可解释人工智能的制度框架[26]。可解释算法虽从理论上能凭借不同的解释路径透视算法黑箱，但现实中，算法依赖于强大硬件的支撑和大量数据的累积，连算法设计者也无法保证完全理解并解释算法运行的全过程，可解释算法只限于乌托邦式的设想。已有案例表明人工智能可以在专业人士都未意识到的情况下精准识别种族身份。2021 年 8 月，有报道称一人工智能模型通过 x 光片和其他医学图像中的数据训练，学会了高精准识别患者种族身份。“令人沮丧的是，这项新研究的开发者无法弄清楚他们的模型究竟如何能够如此准确地检测出患者自我报告的种族。”[27]这意味着，医疗人工智能工具能够检测到患者的种族身份，并有可能在算法设计者或医生完全无意识的情况下将种族作为其输出和决策的一个因素。

(三) 价值偏好的隐性渗透

多年来，科学界普遍持有这样一种信念：研究者应使用不带任何主观偏向的语言对自然界进行描述，避免在科学探究过程中掺杂个人的理论立场或价值判断[28]。但事实上，技术进步一定是在社会中发生作用而并非在真空中[29](157)，我们称之为知识的内容，皆源于并依托于社会群体中的共识。由于各社会团体固有的价值观，事实与价值之间难以实现真正的分离[30]。虽然计算机“被灌输了程序员为达到特定目的而编写的代码”，但程序员所输入的代码和算法并不是脱离伦理的，人类设定的任务也并非价值无涉的，它们恰恰体现了一定的目标和偏好[29](159)。

因此，即便是假设所有的算法开发者都没有歧视意图，未对算法进行积极的人工编辑干预，他们无意识的价值输入也依旧会影响算法的“中立性”。目前，大规模的人工智能系统几乎只在少数科技公司和少数精英大学实验室开发，而在西方，这些实验室的工作人员往往具有几个固定标签：富人、男性、白人。数据显示，2023 年谷歌只有 5.6%的员工是黑人[31]，2022 年 Facebook 女性科技人员占比是 25.8%，黑人员工占比是 4.9%[32]。因此而产生的不平等带有明显的交叉性特征，“这种制度的好处，从利润到效率，主要来自那些已经处于掌权地位的人，他们同样往往是白人、受过教育的男性”[33]。

自然，我们无意评价此类科技公司的招聘规定，许是由于国外阶级固化产生的受教育程度参差导致此类企业即便尽力践行多元、包容的招聘理念，依旧无法缩小现有的就业群体差距，但技术人员因此产生的无意识的和潜在的种族偏见、性别偏见也许会体现在算法的“价值观”中。正如研究所表明的，很少有城市数据集跟踪和分析性别数据，因此，基础设施项目通常不会考虑女性的需求[34]。技术设计之初是为了服务人类，若设计者不对其作批判性验证和思考，他们会在无意中成为社会不平等的推动者。

当然，强调从业者的“去意图”价值偏好并非因此否认以“故意”为导向的情形存在，与传统歧视行为相比，算法歧视因其生成及显化载体呈现出更多无意识性。现实中依旧存在算法设计者在算法设计、运行过程中积极追求并实现区别性结果的情况。例如，在英国，克里斯·阿东(Chris Atton)分析英国民族党(British National Party)的网页后发现，该组织充分利用网页这一选择性媒介将种族主义塑造为对外界压迫的自然反应[35]。

(四) 利益导向的社会生态

意识到算法对社会平等带来的潜在威胁，大量研究试图通过克服人工智能的技术难题而达到实现算法正义的目标。事实上，仅仅在技术层面修正算法虽然能够在一定程度上缓解算法不公正引发的公众质疑和社会矛盾，但无法根除本质问题——“这不会解决社会问题，而是会将它们掩埋在政治正确和经过净化的幌子下”[36]。造成算法歧视的社会根源之一是以逐利性为核心的资本导向的社会生态。为资本逻辑代言的新自由主义长期处于西方经济理论主流位置，主张金融化、全球化、自由化，强调“自由”的根本要求就是免于干涉，国家需让位于市场[37]。新自由主义并非停留在理论层面的纸上谈兵，“作为意识形态、理论流派和制度实践的综合体，新自由主义因契合资本进一步扩张的需求和中心国家时代的利益诉求，一步步成为资本主义发展的新的具体制度”[38]。

就宏观角度来看，利益导向的社会生态“使得财富高度集中于发达国家的富人阶层，金钱也拥有了更大的政治权力，大资本扶持的政治势力总是能够使政策向资本家阶级倾斜”[39]。而国家长期在这种经济背景下运作，并通过法律和政策维持现有社会结构，使等级关系合法化。这也意味着，奉行新自由主义理念和政策的国家，可能难以实现真正意义上的社会平等。长此以往，这种根植于文化认知与社会实践的结构性不平等根本无法消除。

具体到算法层面，利益导向导致适用于人工智能技术的许多伦理原则包含了功利主义导向——“其基本原则对大多数人来说是最好的结果，而这也意味着永远不会寻求以少数群体为中心的解决方案”[40]。值得注意的是，占据支配地位的群体甚至由于利益至上价值的潜移默化而意识不到他们的不公正做法：“公司通常并不自觉意识到它们的政策是如何延续歧视；它们根据通行的标准雇佣员工，仅仅是试图最大化利益。”[41]

三、算法歧视对社会结构的反建构

著名社会学家达伦多夫认为，群体之间的各种利益差异越是相互叠加在一起，比如贫富的差异又叠加上种族的差异，群体之间的冲突就会越激烈[42]。现实中，受到算法歧视不良影响的群体客观上多具有负面交叉性特征，例如老年人、有色人种、贫困、残疾人、女性，由算法歧视所加剧的社会矛盾也会愈加突出。可见，算法歧视与社会发展互施影响，对赖以生成的社会结构具有反建构作用。

马克斯·韦伯采用多向度的社会分层理论来划分社会阶层，并设定三个标准，即经济标准——财富，政治标准——权力，社会标准——声望[43]。具体而言，经济维度强调经济因素在决定个人和群体的社会地位中的重要性，重视个人或群体对经济资源的控制；政治维度包括能够影响社会政策和决策

的能力；社会维度与社会荣誉和声望有关。三者相对独立又相互影响，合力形成制度化的不均等体系。从现实情况来看，算法歧视与失衡的社会结构相互强化，从财富、权力、声望三个角度固化了社会阶层，阻碍合理化社会流动。

(一) 财富——算法歧视加剧贫富差距

经济资源在社会分层中居于首要位置，由此产生的穷人富人之分是公众对于阶级甚至阶层的最普遍印象。阶级代表个人在经济秩序中所处的地位。韦伯强调市场因素在阶级中的重要性，认为市场地位就是阶级地位，只有自由交换和自由流动才能形成阶级[44]。也有观点对经济资源采取更广义的理解，认为其除了市场地位，还包括生产资料的占有或剥削与被剥削以及收入水平[45](9)。广义视角有助于我们多角度理解算法歧视如何固化支配货物或劳动收益的资格或权力，加剧贫富差距，阻碍财富流动。

1. 生产资料的占有

作为生产力中物的因素，生产资料是经济资源的重要组成部分，其与人的关系被视为马克思主义阶级划分的依据。就传统生产资料来说，算法歧视对于少数群体的收入水平、医疗资源、社会福利甚至是犯罪预测等方面的不利影响已被研究证实[46]。这种全方位的不平等反映在经济上，影响着受歧视群体对于技术、知识、土地及自然资源等传统生产资料的占有。此外，随着数字经济时代的加速推进，信息与数据已经成为新的生产资料形式。2020 年《中共中央 国务院关于构建更加完善的要素市场化配置体制机制的意见》明确将数据列为生产要素。不平等的算法决策在招聘、岗位设置等环节的影响，导致女性、年龄较大者、少数族裔等群体难以进入高科技行业，更不必说对数字技术和资源的掌握与占有。

大数据的生产在一定程度上体现了剥削与被剥削。数据本身就是可出售的商品，同时又是能够被用于用户偏好和消费习惯分析的价值待实现的产出品，本质上属于商品资本。“数据是劳动创造的，资本家与劳动者之间存在剥削关系，只不过这种剥削关系因为有娱乐的性质而显得更为隐蔽，但是扼杀不掉劳动对资本的附属关系。”[47]从算法歧视角度而言，公司通过不平等的算法决策支付给受歧视群体低工资、提供职业发展前景低的体力劳动岗位，体现了歧视性算法加剧此类群体受剥削程度的现实。

2. 收入水平

毋庸置疑，算法歧视对于收入的影响具有直观性。研究表明，招聘算法在筛选简历时会对女性、残疾人、特定族裔有偏见，从而限制他们从事高收入职业的机会[48]。相应地，这些低收入群体在信贷和金融服务中也无法获得高阶评分，只能以较高利率获得贷款。此外，历史和现实因素造成的受教育水平差异、医疗资源差异等情况，也可能导致特定群体在学历、身体状况等方面的竞争力下降，间接影响其获得高收益的资格和机会。此类情况无一不限制特定群体的财务灵活性，影响其经济状况和社会阶层位置。

3. 市场地位

韦伯强调市场因素在经济秩序中的重要性，认为任何类型的享用商品、财富、劳动收益的资格或支配权力，都会构成一种特殊的阶级地位[49]。市场地位标准综合了一个人多方面的生活机会和生活状况[45](14)，由市场流动和交换形成的不同利益体包括债权人与债务人、雇主与雇员等。就市场地位而言，算法技术的大规模应用已经极大地限缩了低支配能力群体上升的阶级流动渠道。在数字时代，公司低级员工的工作内容很可能完全由算法管理和匹配，可视化的职业上限使得类似 Charlie Bell 从煎汉堡工到麦当劳 CEO 的职业发展路径在算法时代变得不太可能。以网约车服务为例，一方面，系统权限可及性受限使得网约车司机所能访问的工作信息方式和内容完全由公司决定和提供，这种分流信息层[50]使得司机无法详细了解工作基本过程；另一方面，公司利益最大化的高效率调度安排使得司机难以有时间和精力去接触基本驾驶服务以外的其他工作信息，基本杜绝了司机向上岗位流动的可

能性。数字时代的工作趋向如此，更不必说这些匹配岗位的算法还带着歧视，因为算法歧视更加凸显了受歧视群体在“占有财产、占有某种商品、占有某种信息、占有某种机会、占有某种市场的能力”[45](9)方面的短板，制约了这一群体市场地位的提升。

(二) 权力——算法歧视导致社会权力失衡

1. 算法歧视损害法理型权威

韦伯认为，政治体制造成不同政治分层的群体拥有不同的权力，处于政治分层顶端的群体可以通过行使手中权力来影响市场分配[51]。达伦多夫认为，社会分层的不平等体系是社会权力结构的派生物，在合乎法律和期望的条件下，权力便转变为权威。于是，一些社会角色就用一种合法的权力去统治和强制那些处在从属地位上的社会角色[45](57)。合法性是促使一些人服从某种命令的动机，关系到政治行为能否得到人民的信任、认同和服从[52]。随着以规范和规则为基础的秩序之治深入人心，建立在民主与法治之上的法理型政党权威取代传统魅力型权威和绩效型权威[53]，成为现代社会政治合法性的基础。

在法理型权威的价值体系中, 法律是最终的合法性来源[54]。法理型权威是基于法律正当性和程序合理性的权威形式，支撑着政治决策的合法性、正当性与民主性。当前，算法已被广泛应用于政府事务中，通过大数据描述人类行为和社会现象，政府部门参照算法决策结果做出公共决策。带有歧视导向的算法在社会决策应用过程中损害正当程序的中立性、公开性、论证性、公正性[55]。从权威来源角度而言，算法歧视所做出的公共政策影响公众对权力的自愿服从，损害法理型权威。

2. 算法歧视影响政治参与的平等性

从现代民主视角来看，普遍的选举权和被选举权是实现政治参与平等的重要保障。当前，在选民注册和投票过程中，国外部分地区使用面部识别技术来验证选民身份，而这却因为算法歧视的存在导致选举权和被选举权的平等实现受到影响。在现实中，人脸识别技术多次被证实识别黑人的准确性较低，可能因此影响部分选民投票权的实现[56]。同时，算法已被认为正通过选民操纵等方式异化西方选举政治生态[57]，若算法在政治广告投放或候选人推广方面存在偏见，通过降低可见性或限制推广人群等方式影响少数群体在政治领域的代表性，将进一步加深政治分层。

现实中，算法应用于分析社区需求以进行资源分配，而由偏差性数据所产生的歧视性结果会因此影响政策的公平性与合理性。波士顿曾使用“Street Bump”应用程序来检测城市道路上的坑洼并予以修缮。驾驶员通常将智能手机放置在汽车内部，当应用程序检测到汽车行驶时的颠簸情况，该颠簸数据就会被用来确定道路存在坑洼之处[58]。该 App 被实际应用后发现，老年人聚集区及较为贫困地区因汽车数量较少、智能手机普及率低以及 App 应用范围较小等原因而呈现出路况良好的假象，这无疑加剧了这些地区公共资源获取的不平等性。此外，互联网平台被证实通过算法过滤少数群体用户的声音[59]，少数群体的需求和声音不易被重视，在政策制定和执行过程中的代表性缺失导致少数群体在社会政治结构中的位置更加边缘化。

(三) 声望——算法歧视加剧社会偏见

声望体现的是人与人之间的主观评价，在社会互动和公众认可的过程中被建构，与文化、传统和社会价值观紧密相关。韦伯将由受到同样的肯定或否定社会声望评价的人构成的群体称作身份群体(status group)。从声望角度而言，广泛存在的算法歧视固化了社会评价标准，在潜移默化中加剧了社会偏见。

1. 算法歧视塑造偏见型社会认知

米歇尔·福柯的话语权力理论认为，话语不仅仅是言语或文本，而是一种实践，权力通过话语实践来行使，塑造了我们对世界的理解及采取行动的方式。“任何话语场域都是重要的社会化场所，场域中话语权的拥有与改变虽然是各种因素的复杂聚合，但权力占据者却具有明显的优势支配权。”[60]

算法歧视事实上就是一种话语实践，利用决策支配权力，通过编码来实施特定的知识形式和偏见。

算法，尤其是公权力使用的算法决策系统已被证实加剧了对某些社会群体的不公正对待。举例来说，美国的有色人种社区因算法预测性警务系统受到过度监控，数据显示，黑人因吸毒被预测为治安目标的概率是白人的两倍，其他有色人种也高出白人 1.5 倍，尽管现实中各族群的吸毒率基本相当[46]。由此生成的“有色人种需加强监控”的话语实践塑造了公众的认知。如果说前数字时代公众对于特定族裔的负面印象和评价因“消除一切形式种族歧视”“人人生而平等”的明文规定而居于“心证”状态，那算法时代的这种带有高科技、科学性标签的显性歧视无疑给这种“心证”提供了“力证”，与社会达尔文主义利用“适者生存”粉饰种族歧视异曲同工。这种歧视性算法运用在公共决策中，更会因表面上的公权力背书而更具可信度。长此以往，对于特定群体的刻板印象会随社会实践传播，并在偶发性事件中因被“证实”而加深。

2. 算法歧视限制少数群体话语空间

话语通过规定什么可以被说、谁可以说话、在什么场合说话等，对知识和社会实践施加控制[60]。社交媒体近年来推广使用内容过滤、推荐系统及自动审查机制来处理海量信息。若是算法基于偏见数据集训练，可能导致某些群体的内容被系统性地降低可见性。缺乏文化代表性和广泛性的算法无法正确区分少数群体的文化和语境差异，导致其言论被错误过滤，限制了少数群体的话语空间。有研究指出，谷歌自然语言处理模型 NLP 数据集已被广泛过滤，删除了黑人和西班牙裔作者、同性恋者及其他少数群体的源数据，研究人员还发现非裔美式英语和西班牙裔英语受到黑名单过滤的影响尤为严重[59]。此外，搜索引擎可通过算法确定网页的排名和可见性，而对比看来，少数族裔的信息在互联网上的可见性较低[61]。

在稳定的社会和按照常规运作的社会里，社会分层是一种社会结构或高低不同的位置结构，至于谁进入哪一种位置，是由社会流动决定的[45](9)。在任何社会中，社会流动被认为能减少阶级冲突，是社会稳定的“安全阀”[62]，社会资源的分配、交换和转移都通过社会流动来实现。新韦伯主义者帕金认为，向上社会流动(upward social mobility)为底层中最有能力和最有抱负的大量成员提供了一条逃离路径，从而缓解了不平等产生的一些紧张关系[63]。总的来说，歧视性算法从构成社会阶层的经济、政治、社会三个维度多层次全方位恶化了少数群体在社会结构中的不利处境，此类群体通过正常渠道上升流动、获取更多社会资源的可能性愈发小，本就不平等的社会结构正随着歧视性算法的普及而愈发失去平衡。

四、 算法正义的建构标准

在算法歧视的治理问题上，许多学者从包括算法技术优化、法律规制、从业人员道德培训等在内的多个角度进行了具体讨论，以实现算法正义，达成“人人在尊严和权利上一律平等”的目标。但这些建议得以顺利实施的大前提是必须明确理想状态的算法应根据何种标准来设计和实施，我们所追求的算法正义的要求和标准为何。

正义是个相当宽泛的概念，不论是柏拉图的“各司其职，各守其序”，还是休谟的“公共福利是唯一源泉”，均体现出不同思想家对于“社会制度的首要德性”的理解和追求。在此基础上，不同正义观指引下的学者们所坚持的理想算法风格迥异，对消除算法歧视的实践设想也就千差万别。因此，只有在对算法正义的建构标准做出界定的基础上再探索算法正义的实现路径，才能在理论上经得住推敲，在应用中实现理想目标。举例来说，如果要设计一款向游客推荐安全舒适的居住地点的旅游 App，软件开发者是应无偏差地推荐游客根据 App 到达犯罪率相对高的街道和社区，还是应根据犯罪率经验

数据指引游客到更安全的地方，同时陷入导致经济不发达地、老龄人口聚居地住宿行业因此发展停滞的尴尬境地？在这个例子中，采取何种标准界定算法正义，并以此进一步指引设计者在实践中选择何种算法是问题的关键。

(一) 分配正义——受保护特征应被纳入算法决策

旅游 App 设计的核心问题在于，决策中是否应该将受保护特征作为考量因素，即是否应当根据受保护特征对算法决策结果加以调整。进一步而言，这本质上也是算法决策的形式平等与实质平等的问题。这一权衡并非仅仅局限于算法歧视领域，而是自反歧视共识形成之时就备受关注，美国法院近年来关于大学招生政策的论辩就是这一问题的现实写照。为了促成种族多元化，美国一些学校规定了招收部分族裔学生的比例，引发了公众关于“逆向歧视”的质疑。有观点认为，平权保护会给某些族裔打上耻辱的烙印[64]。美国联邦最高法院于 2023 年 6 月 29 日裁定哈佛大学和北卡罗来纳大学考虑种族因素的招生政策违宪，从而事实上禁止美国高校在招生过程中将种族作为考量因素②。我们对此判决结果持反对态度，我们认为，应当在决策中承认种族差异，并将这一受保护特征作为决策的有限考量因素。

被誉为 20 世纪西方最重要的政治哲学家之一的约翰·罗尔斯认为，“正义的首要主题是社会的基本结构，或更准确地说，是社会主要制度分配基本权利和义务，决定由社会合作产生的利益之划分的方式”[65](6)。他提出的“公平正义”理论认为，只要满足两个正义原则，不平等并不等于不公平。第一，每个人应享有平等的基本自由权；第二，社会和经济不平等必须有利于最少受惠者(差别原则)，并在机会平等的前提下向所有人开放职位和地位(机会平等原则)[65](237)。根据第一个平等自由原则，每个人都应享有最广泛的平等基本自由权，这要求算法决策系统在设计和执行时，必须确保不侵犯个体的基本自由。因此，反对算法歧视自然成为该原则在算法时代应用的应有之义。

仅仅“反对歧视”不足以解决社会不平等问题，多元平等的核心在于“回归主流”(mainstreaming)，即不仅宣布歧视非法，更要消除制度性障碍，营造公平竞争的环境[66]。以种族问题为例，虽然自由派强调起点平等，甚至部分美国黑人也认为应靠个人努力而非特殊照顾，但现实中，起点并不真正平等。个人的自然禀赋和社会条件差异共同影响其资源获取与能力发展[67](111)。贫困家庭出身的少数族裔学生往往难以获得优质教育与医疗，在大学录取、就业乃至算法评分所涉及的贷款、福利、量刑等环节中也将处于不利地位，形成结构性不平等的连锁效应。

罗尔斯的第二原则通过机会平等与差别原则回应了现实中的不平等问题，其核心在于对最少受惠者的合理倾斜，体现了通过补偿性再分配实现实质平等的理念。在此问题上，欧洲人权法院也为我们提供了相似的思路。欧洲人权法院认为，《欧洲人权公约》第 14 条③并非禁止公约所承认的权利和自由行使过程中的任何差别待遇。如果“特定一些法律上的不平等仅仅趋向于纠正事实上的不平等”[68]，那这种差别待遇就是允许存在的。从这个意义上而言，在算法决策中纳入受保护特征有利于实现“各种地位不仅要在一种形式的意义上开放，而且应使所有人都有平等的机会达到它们”[67](111)的设想，符合罗尔斯的第二个原则。

考量受保护特征被认为是对最少受惠者不公正待遇的“惭愧”补偿措施，同时也是正视不平等现实、推动社会合作的平衡手段。以美国种族歧视问题为例，有观点认为，“仅仅因为他们的祖先而对公民进行区分，对于一个制度建立在平等原则之上的自由人民来说，本质上是令人憎恶的”④。而现实情况是，不论是“他们的祖先”还是现在的“他们”，都尚未触及本质上的平等。Statista 全球统计数据库 2022 年 9 月 30 日发布的报告显示，2021 年，美国有 19.5%的非洲裔生活在贫困线以下，而白人的贫困率仅为 8.2%[69]。罗尔斯强调互利互惠的合作体系，其正义原则所要求的不是对多数群体的剥夺，而是要求较有利者做出一种捐助和贡献。这是因为，较有利者的福祉本就依赖于社会合作，唯

有在公平条件下，较不利者才可能自愿参与其中[67](126)。

(二) 关系正义——算法决策应对弱势群体予以优待

罗尔斯的分配正义论着重从物品、资源的占有和分配去探讨现实中的不公平问题，影响着世界范围的规范性学说和政治实践，但这种停留在社会基本结构下的正义观似乎无法完全解决存在于各种社会关系中的不公平问题。例如，荧幕中有色人种难以扮演正面角色或犯罪预测系统对有色人种的高风险判断，这些现象所反映的不是物质分配不公，而是文化与象征层面的不平等。针对分配正义存在的理论与实践不足，以艾丽斯·M. 杨为代表的关系正义学者关注成员自我决定和自我发展的实质机会与能力，注重从社会过程和社会关系的视角来理解正义。

1. 尊重群体多元并反对多样性“武器化”

法律、道德等社会规则承诺维护群体之间的平等，公开的歧视已经为正式的社会规则所禁止。但如前文所述，算法歧视还受到从业人员价值偏好及算法的高度专业壁垒等因素的影响，这意味着消除歧视还应从非正式场合，特别是从交互习惯、无意识假定和刻板印象等方面入手。杨认为，正义的范围涵盖了支持或破坏压迫的所有社会过程，包括文化。她提出要肯定自身及其他群体的异质性和多元性，“为了让人们与身边那些被他们认为不同的他人融洽相处，或许有必要让他们先学会与包含在自身之内的异质性融洽相处。”[70](186)

就消除算法歧视而言，多元(diversity)被认为有利于促进不同群体成员之间的思想交流、相互理解。如前文所述，人工智能从业者的单一性背景是导致算法歧视的重要因素，因而需要在国家和社会层面倡导多样化背景以实现多元目标。但应警惕的是，科技公司出现了将多样性语言“武器化”的倾向。谷歌于 2019 年宣布成立先进技术外部咨询委员会(ATEAC)，旨在对其人工智能技术进行伦理审查。该委员会成员之一是传统基金会主席 Kay Coles James，一位坚定的反移民者、反 LGBT 人士。谷歌声称这一任命是为了体现“思想多元性”，尊重保守派声音。谷歌员工对此表示集体反对，超过 2 300 名员工签署了一份要求公司将 James 从专家组中除名的请愿书。请愿书指出，“任命 James 加入 ATEAC，等于谷歌强化并认可了她的观点，暗示她的观点值得被纳入决策。这令人无法接受”[71]。

在美国，国籍、移民身份均是受法律保护的个体属性。更重要的是，移民问题与种族问题高度关联，虽不能将二者完全等同，但美国严格的移民政策是以种族为基础的，种族主义贯穿着美国移民政策及对待移民的态度[72]。谷歌在其人工智能道德审查部门吸纳与公司甚至是全美社会既定价值观相反的人士，系将“多元性”这一目标转化为加剧算法歧视的武器。试想，若国内一家大型 AI 企业在其负责审查算法道德伦理的部门吸纳并重用了挑拨性别对立的专家，并借口这是容纳不同观点的多元化操作，也无疑是不具说服力的。

2. 正视基于群体差异的“特殊优待”

杨认为，作为所有群体共同参与、相互包容的平等，有时需要对受压迫或弱势群体予以特殊对待[70](192)。她提出，超越群体差异是一种追求同化的理想，坚持群体差异的解放政治涉及平等意义的重新概念化[70](192)。由此出发，应正视受压迫或弱势群体因为歧视性算法而受到的不平等待遇，并对其予以“优待”。当前，针对在人工智能领域给予少数群体“优待”并提高其群体代表性和话语权的政策尝试，遭到了部分既得利益者的反对。反对者认为，提高算法公平性和包容性的核心在于提高“认知多元性”，即人们思考和理解世界的个体差异与其他群体差异是相同的——“十几个白人男性，只要他们不是在同一个家庭长大，就不被认为其有相同的想法，也就可以被认为是多元化的”[73]。反对者还认为，提高特定群体在科技行业就业的比重相当于承认他们的“不合群”，这反而是一种排他行为，不利于消除歧视。换句话说，反对者认为，即使科技公司所有员工都为白人男性，其思想和认知也会因为各自生活环境和经历的差异而呈现“多元性”，因而无需通过平衡其他群体在雇佣中的比重来达到人工智能输出观点多元化的目标。

我们认为，上述反对论既忽视了造成社会歧视的历史因素，也忽略了维持社会歧视的现实情况，更是无视了不同群体之间在权力和话语方面的差异。杨批判了这种超越群体差异的同化主义理想，认为忽略群体差异会在三个方面加重社会压迫。首先，无视群体差异会强化某些群体的弱势地位，因为这些群体在经验、文化、社会化能力方面有别于特权群体。在同化策略中，衡量所有人的标准由特权群体制定，而受压迫群体在被这些看似“中立”的规则衡量时往往处于劣势。其次，同化理想预设了普遍善的人性，但现实中根本不存在能够脱离社会情境和经验的观点，那么支配性群体的经验就会被视为界定人性的普遍规范。少数群体因为和这些普遍规范存在差异而被视为特殊的、非正常的，这使得文化帝国主义有了“充分”的理由。最后，偏离所谓中立标准的群体将遭到诋毁，这往往导致这些群体成员的内在化贬值[70](200−201)。杨指出，“在那些存在着社会群体差异，有人特权加身、有人遭受压迫的地方，社会正义要想打破压迫，就必须明确地承认、关注这些群体差异”[70](2)。

3. 扩大少数群体算法决策话语权

参与式民主是社会正义的一个要素和前提[70](223)。杨指出，在缺少哲人王的情况下，做出一项正义决定的唯一基础是能够真正促使所有需求和观点得到自由表达。“所有人都应拥有权利和机会去参与制度的审议和制定，他们的行为将会对这些制度有所助益。”[70](110)在算法问题上，就是要提高算法决策的群体参与平等性，扩大少数群体决策的代表性和话语权。

现实中因决策过程缺少代表性导致决策结果不公正的例子屡见不鲜。一项针对北美一大型零售连锁店的研究显示，在根据客观成就获得的“绩效”评分上，女性员工高于男性员工，但在职位晋升中，领导层主观考量“潜力”因素占比更大[74]。也就是说，即便算法计算生成的绩效评分结果客观中立，公司也不会参照决策结果反而是根据“男性更有潜力”的主观印象决定晋升人选。有科技公司总裁甚至直言，女性的聘用和晋升是基于过去的成就，而男性的聘用和晋升则是基于未来的潜力[75]。若该公司的晋升机制能有效地承认和代表女性群体的声音，为她们从“我想要”变成“我有权”提供更多话语权保障，那么，相信这种“更有潜力”的刻板印象和“双标”的晋升考量会在民主声音的表达和监督中逐步收敛乃至消除。

五、迈向算法正义的技术、法律、伦理三元协同治理

在学理层面明晰了算法正义的建构标准之后，需进一步探索的是如何于算法实践中，切实嵌入算法正义的模式架构。事实上，当下 AI 的“价值取向”问题已引发各界关注。例如，欧盟 2024 年发布的《人工智能法案》(*Artificial Intelligence Act*)规定，人工智能及其监管框架的发展必须符合《欧洲联盟条约》第 2 条所载的欧盟价值观。我国聚焦算法治理议题的《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《算法指导意见》)也规定要弘扬社会主义核心价值观，在算法应用中坚持正确政治方向、舆论导向、价值取向。然而，如何在实践中促进不具有约束力的倡导性目标在算法实际应用中得以落地，仍需进一步细化设计。有学者总结，人工智能伦理规范的非强制性、抽象性、模糊性，分散、混乱与叠床架屋，自愿遵守的动力不足，合规悖论，社会系统论困境等共同导致了人工智能伦理规范的“实施赤字”[76]。有鉴于此，本文在厘清算法正义的建构标准之后，将着力点置于探索将这一理想化的“目标导向”融入具体制度设计之上。

(一) 构建以算法区分为基础的法律规制模式

现实中，算法技术已广泛渗透于多元领域，不同类型算法因设计逻辑、功能目标和风险特性各异，其规制路径和治理成本必然存在差异。我国《算法指导意见》提出要“健全算法分级分类体系”，其核心意图即在于由此实现监管资源的合理分配和治理效能的最大化。尽管政策方向已明确，但具体如

何构建和实施“分级分类体系”，以及如何在实践中推动管理工作的系统化与精细化，尚无进一步规范和指引。面对算法规制的复杂难题，Omer Tene 和 Jules Polonetsky 在 2017 年提出基于“政策中立算法”和“政策导向算法”的区分理论[36]，认为应针对应用主体和决策辐射范围区别设计不同应用场景下的价值融合方式，以此引导算法在可控、安全、人权框架下健康发展。政策中立算法强调客观性和中立性，要求算法在设计和实施过程中不偏袒任何特定政治、社会或意识形态立场，公正、客观地处理数据和提供决策。而政策导向算法则是指算法在被设计和实施过程中被有意地考虑特定政策目标和社会需求，旨在利用数据和技术的力量来优化政策的实施和决策制定过程。我们认为，基于算法区分理论，可以进一步构建以“算”前设计、“算”后调整为核心的法律规制路径，从而进行算法规制的分级分类，实现算法的针对性管理与差异性监管。

1. “算”前设计的法律规制：基于场景理论的算法决策应用

Omer Tene 和 Jules Polonetsky 虽对政策中立算法和政策导向算法进行了理论区分，但未对两类算法的实际应用做进一步配套设计。推进算法正义的目标在于实现公共治理领域的社会公平，这一过程不能忽视自动化应用所带来的效能提升。欧盟《人工智能法案》出台以来饱受“超前监管”的质疑，被认为将使 AI 企业承担过高成本，影响人工智能开发进程[77]。AI 治理的关键在于如何在促进技术创新与实现社会公平之间找到动态平衡，以推动二者的协调发展。

为实现技术效率和社会公平的“两手抓”，我们主张参照场景理论，以情境为导向考察自动化应用的风险水平，基于算法决策的主体和辐射范围来区别适用两类算法形式，从而实现符合技术发展实际的法律规制目标。具体而言：第一，尊重政策中立的法律规制。若算法是用于个人、企业等私主体且决策事由是类似于企业自主经营权的内部决策事项，或适用于虽然具有外部面向性但其目标追求具有高度客观性的决策事项(如算法分析数据做天气预报)，则适用不受特定政策目标影响的政策中立算法，以实现算法决策的效率、效益和客观性。第二，适用政策导向的法律规制。若算法作为公共决策的标准和依据，特别是应用于与公众政策、社会福利相关的具有广泛公众影响性的场景中，则应根据不同应用情境适用特定政策目标(如性别平等、种族平等)的政策导向算法，即在算法设计时综合考量分配正义与关系正义的共同要求，将受保护特征纳入算法决策，并在考量群体差异的基础上对弱势群体予以优待。当然，算法场景或情境的区分并不是非此即彼的，往往还需要考虑那些具有复杂特征的算法决策场景。例如，大型科技公司、数字平台普遍通过制定行业规则来治理平台市场环境，这种准公共空间的特性推动此类“私主体”具备了一定的“准公权力”。因此，在场景理论应用中不能单以决策主体来判断，重点是依据决策的数据特质、目标群体、结果辐射范围等因素加以考量。对于追求客观中立的分析研究(如科学研究、医疗诊断)、不特定于政策目标的通用决策支持(如股票分析、天气预报等)、类自主经营的商业市场运营(如消费者行为预测)等，应秉持政策中立的算法立场施以宽松的法律规制，以提升算法决策的社会效益。对于具有公共影响性的公权力或准公权力算法决策，如医疗资源配置、警力部署优化、公共资金投入分配等事项，则应秉持政策导向的算法立场，通过严格的法律规制来确保社会平等目标的实现。目前，我国在算法规制问题上尚未确立区分框架，这不仅无法充分释放算法技术的潜力，也难以有效回应社会关切。只有构建场景化的算法规制体系，才能实现技术进步与社会公平的有机结合，避免算法成为社会不公的放大器。

2.“算”后调整的法律机制：决策结果的审查与修正

政策中立算法追求决策效率和结果客观性，但正如前文所述，由于历史数据偏差、算法高度技术性及科技人员的价值渗透等因素的综合作用，算法不可避免地会蕴含已有的社会偏见和不平等，使得“客观”的结果“不正义”。因此，亟须建立决策主体伦理审查和监管制度，常态化监测、评估算法决策结果，修正偏差性、歧视性决策。

第一，对算法决策结果进行伦理审查。《人工智能法案》第 14 条提出了自然人监督者介入 AI 系

统的要求，监督者具备介入系统运行和中断系统的权限，以此防范因合理预见的误用而可能引发的风险。这一规定强调了自然人在价值判断与伦理抉择上的不可替代性。当前，我国已推动建立单位科技伦理委员会制度。2022 年 3 月发布的《关于加强科技伦理治理的意见》规定，从事生命科学、医学、人工智能等科技活动的单位，如果研究内容涉及科技伦理敏感领域，应设立科技伦理(审查)委员会。“算”后调整机制可以依托伦理审查委员会发挥作用：任何参照政策中立算法输出的结果，都应由伦理审查委员会遵循公认的运行机制进行审查，对于不符合伦理的决策依据算法正义来进行修正。例如，针对算法根据物流、居民预期需求等因素做出的大型超市在偏远少数民族聚集地区不提供快速送货的决定，如果伦理审查委员会认为这种以历史数据和利润最大化为导向的决策带有社会偏见，应修正算法结果，选择对此地区予以区别对待，提供相应服务以促进社会福利。

第二，算法决策的透明度义务。算法决策应提高透明度和可解释性已经成为各国实践共识，我国《个人信息保护法》第 24 条、美国《算法正义和互联网平台透明度法案》(*Algorithmic Justice and Online Platform Transparency Act*)均对此有所规定。在算法区分的实践框架下，由于政策导向算法的决策结果通常具有社会性和公共性，为消除“权威从个人转向由算法构成的网络”[81]的担忧，算法应用的公权力主体所扮演的角色和其所推进的伦理价值的重要性日益彰显。因社会治理的需要，公权力部门经常需要同掌握大数据的数字平台合作。在这一背景下，算法外包，尤其是预测型算法的这种权力“外包”挑战了正当程序和原则。公开、透明、公正参与被公认为现代正当法律程序不可或缺的内容，在此基础上，用户有权知道影响个人权益的算法决策背后是何种价值的推动，有权关切自己所接触的是如何进行过“修剪”和编辑的世界。在先前提及的旅游 App 推荐住宿地点的例子中，由于涉及社会安全和稳定的公共利益，科技公司应适用政策导向算法，在代码设计中融入平等的价值理念，在帮游客推荐居住地点时也提供位于高犯罪率街道的酒店，但应以明显的方式提醒游客，其所展示的搜索结果是经过设计和优化的。也就是说，相较于政策中立算法而言，采用政策导向算法的主体承担着展示更多“透明度”以实现程序正义及算法问责的义务。据此，应对采用政策导向算法的主体施加强制影响评估和强制披露的双重义务，即要求此类公共(准公共)主体必须进行算法影响评估，预测并减轻对社会的潜在不利影响，尤其是对弱势群体的影响。在此基础上，要求此类主体将数据处理、算法目标、理念导向、算法结果、合规情况即评估情况对社会公开，提高算法决策的公共信任度。

(二) 建立“科技向善”导向的人工智能全周期治理机制

人工智能的潜在风险贯穿其研发、应用和退出的各个阶段，各国因此愈发重视全周期治理。美国白宫 2022 年发布的《人工智能权利法案蓝图》提出对自动化系统的设计、使用和部署进行全生命周期的指导；欧盟《人工智能法案》开宗明义，强调“针对 AI 系统的开发、投放市场、投入使用及应用”建立统一的法律框架。在实现算法正义的路径中，同样需要在算法区分的基础上注重全周期治理，以实现“科技向善”的应用目标。

在算法概念化阶段，核心是确立算法正义的思维框架。当前人工智能监管思维呈现出工具理性导向，重点关注算法部署后的结果评测与损害救济。我国《算法指导意见》和《生成式人工智能服务管理暂行办法》均对算法的前端价值嵌入机制有所忽视，使得算法系统在初始架构阶段面临价值缺位的潜在风险。对此，应建立跨学科的价值协商机制，即在算法需求定义阶段，在通过算法目标区分出政策中立与政策导向算法的基础上，组建由技术开发者、法律学者、伦理学家、社会学家及利益相关群体代表构成的委员会，根据算法区分将社会公平诉求转化为可量化的约束条件。而且，为避免算法从业者的价值渗透，政府应在学校教育和从业培训中强制性推广关于文化敏感性和包容性的课程培训，科技公司也应扩大弱势群体在此行业的雇佣比例。

第二，在算法测试和部署阶段，监控和调整算法输出以响应不同群体的需求和反馈至关重要。算法测算时应包括多元化的用户群体，尤其须增强弱势群体的代表性和话语权，并重视收集和评估此类

群体的反馈。我国《算法指导意见》明确，应积极开展算法安全评估。针对这一要求，应引入独立、公正的第三方机构进行持续的公平性评估。评估机构应具备公信力，独立于算法开发者和政府，重点评估算法是否存在偏见或对特定群体的不公正影响。若算法符合标准并投入应用，评估报告应随之公开，确保评估过程与结果的透明度。

第三，监管和问责阶段。《全球人工智能治理倡议》倡导打造可审核、可监督、可追溯、可信赖的人工智能技术。对此，应通过法律明确算法使用中的责任归属和责任追溯机制，同时建立配套追溯体系，包括记录和存档算法的运行日志、数据来源、决策过程等关键信息。立法应要求各方按照标准化程序保存这些数据，以便在发生问题时追溯到具体的责任环节。针对算法结果的反馈与申诉，《生成式人工智能服务管理暂行办法》第十五条提到服务提供者应建立投诉、举报机制，但并未明确规定处理时限和具体操作流程。对此，应通过法规明确投诉与申诉的处理时限、反馈机制及处理流程，确保每一项投诉都能得到及时响应和合规处理。而且，需明确涉及不公平算法结果反馈的优先级，并对整改结果进行公示，以增强透明度和公信力。

六、结语

算法歧视并非由技术发展带来的前所未有的新问题，而是长期存在的社会歧视问题在算法时代的映射。长远来看，仅靠调整算法、更新系统无法消除社会不平等的深层次根源，反而是将已有的社会矛盾隐藏在二进制代码的面具之下。要实现算法公平，从科技层面精进人工智能系统、优化生成环境、提高算法性能自然重要，但若单一地注重技术系统，不考虑其所赖以产生和运行的社会背景，无疑是舍本逐末。仅当一个多元文化规范、规则、法律和意识形态体系占据主流地位，而且在整个群体的社会认知和互动层面上被积极施行和共享的时候[41](26)，社会不平等才会消失。也就是说，算法歧视问题的解决，要求我们从技术、伦理和法律端共同发力。

注释：

① 本文所述的算法歧视指发生在自动化系统中的基于受保护特征的歧视。详细定义参照美国《人工智能权利法案蓝图》：算法歧视，是指自动化系统在运行过程中，因个人的种族、肤色、族裔、性别、宗教、年龄、国籍、残障、退伍军人身份、遗传信息，或任何其他受法律保护的身份类别，而造成无正当理由的差别对待或不利影响之情形。https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

② Students for Fair Admissions v. Harvard, 600 U.S. 181 (2023).

③ 《欧洲人权公约》第 14 条规定：对本公约所规定的任何权利和自由的享有应当得到保障，不应因任何理由比如性别、种族、肤色、语言、宗教、政治或其他观点、民族或社会出身、与某一少数民族的联系、财产、出生或其他情况等而受到歧视。

④ Rice v. Cayetano, 528 U. S. 495, 517 (2000).

参考文献：

[1] REED L, BOYD D. Who controls the public sphere in an era of algorithms? Questions and assumptions [EB/OL]. (2016−05−13) [2024−08−08]. https://datasociety.net/pubs/ap/QuestionsAssumptions_background-primer_2016.pdf.

[2] GUTERRES A. Secretary-general’s remarks at the women’s civil society town hall [EB/OL]. (2023−03−13) [2024−05−04]. https://www.un.org/sg/en/content/sg/statement/2023-03-13/secretary-generals-remarks-the-womens-civil-society-town-hall-delivered.

[3] 王凡, 苗子. 育儿补贴丑闻之下 荷兰内阁集体辞职[EB/OL]. (2023−12−28) [2024−05−04]. https://p.dw.com/p/3nykv.

[4] 习近平. 加强领导做好规划明确任务夯实基础 推动我国新一代人工智能健康发展[N]. 人民日报, 2018−11−01(1).
[5] 徐琳. 人工智能推算技术中的平等权问题之探讨[J]. 法学评论, 2019, 37(3): 152−161.
[6] 江必新, 刘倬全. 论数字伦理体系的建构[J]. 中南大学学报(社会科学版), 2024, 30(1): 38−49.
[7] 丁晓东. 论算法的法律规制[J]. 中国社会科学, 2020(12): 138−159.
[8] 李成. 人工智能歧视的法律治理[J]. 中国法学, 2021(2): 127−147.
[9] 张凌寒. 共享经济平台用工中的性别不平等及其法律应对[J]. 苏州大学学报, 2021, 42(1): 84−94.
[10] NOBLE S U. Algorithms of oppression: How search engines reinforce racism[M]. New York: New York University Press, 2018: 1−14.
[11] BENJAMIN R. Race after technology: Abolitionist tools for the new Jim code[M]. Cambridge: Polity Press, 2019: 41−46.
[12] AMINI A, SOLEIMANY A P, SCHWARTING W, et al. Uncovering and mitigating algorithmic bias through learned latent structure[C]// CONITZER V, HADFIELD G, et al. Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. New York: Association for Computing Machinery, 2019: 289−295.
[13] WATSON-DANIELS J. Algorithmic fairness and color-blind racism: Navigating the intersection [EB/OL]. (2024−02−12) [2024−03−31]. https://arxiv.org/abs/2402.07778.
[14] BROOCKMAN D E, FERENSTEIN G, MALHOTRA N. Predispositions and the political behavior of American economic elites: Evidence from technology entrepreneurs[J]. American Journal of Political Science, 2019, 63(1): 212−233.
[15] 王虎学. “人之谜”的哲学自觉与解答[N]. 光明日报, 2020−07−06(15).
[16] HARDING X. Breaking bias: How algorithmic racism goes unaddressed[EB/OL]. (2022−02−11) [2024−05−04]. https://foundation.mozilla.org/en/blog/breaking-bias-how-algorithmic-racism-goes-unaddressed/.
[17] 邢怀滨. 社会建构论的技术观[M]. 沈阳: 东北大学出版社, 2005: 46.
[18] 张玉宏, 秦志光, 肖乐. 大数据算法的歧视本质[J]. 自然辩证法研究, 2017, 33(5): 81−86.
[19] 孙建丽. 算法自动化决策风险的法律规制研究[J]. 法治研究, 2019(4): 108−117.
[20] BARRETT L. Deconstructing data mining: Protecting privacy and civil liberties in automated decision-making[J]. Georgetown Law Technology Review, 2016, 1(1): 153−159.
[21] YU P K. The algorithmic divide and equality in the age of artificial intelligence[J]. Florida Law Review, 2020, 72(2): 331−389.
[22] 李晓辉. 算法商业秘密与算法正义[J]. 比较法研究, 2021(3): 105−121.
[23] 王怀勇, 邓若翰. 算法行政: 现实挑战与法律应对[J]. 行政法学研究, 2022(4): 104−118.
[24] CAPLAN R, BOYD D. Who controls the public sphere in an era of algorithms? Mediation, automation, power[EB/OL]. (2016−05−13) [2024−06−13]. https://datasociety.net/pubs/ap/MediationAutomationPower_2016.pdf.
[25] 张凌寒. 商业自动化决策的算法解释权研究[J]. 法律科学, 2018, 36(3): 65−74.
[26] 杨志航. 算法透明实现的另一种可能: 可解释人工智能[J]. 行政法学研究, 2024(3): 154−163.
[27] SIMONITE T. These algorithms look at X-rays—and somehow detect your race [EB/OL]. (2021−08−05) [2024−04−12]. https://www.wired.com/story/these-algorithms-look-x-rays-detect-your-race/.
[28] 林德宏. 关于社会对技术的必要约束: 评技术价值中立论与价值自主论[J]. 东南大学学报(哲学社会科学版), 2000, 2(3): 15−19.
[29] 马长山. 迈向数字社会的法律[M]. 北京: 法律出版社, 2021.
[30] 肯尼斯·J. 格根, 王波. 历史与关系: 社会建构论的社会建构[J]. 国外社会科学, 2016(5): 135−139.
[31] Google diversity annual report 2023[EB/OL]. (2023−06−30) [2024−04−15]. https://static.googleusercontent.com/ media/about.google/zh-CN//belonging/diversity-annual-report/2023/static/pdfs/google_2023_diversity_annual_report.pdf?cachebust=2943cac.
[32] WILLIAMS M. Embracing change through inclusion: Meta’s 2022 diversity report[EB/OL]. (2022−07−19) [2024−08−15]. https://about.fb.com/news/2022/07/metas-diversity-report-2022/.
[33] WEST S M, WHITTAKER M, CRAWFORD K. Discriminating systems: Gender, race, and power in AI[EB/OL]. (2019−04−01) [2024−04−21]. https://ainowinstitute.org/discriminatingsystems.html.
[34] FLEMING A, TRANOVICH A. Why aren’t we designing cities that work for women, not just men?[N]. The Guardian, 2016−10−13(26).
[35] 王伟. 欧美数字种族主义的演变、机制与根源[J]. 民族研究, 2023(2): 29−44.

[36] TENE O, POLONETSKY J. Taming the golem: Challenges of ethical algorithmic decision-making[J]. North Carolina Journal of Law & Technology, 2018, 19(1): 125−164.
[37] 姜静. 从“自由”概念的历史演变反思新自由主义治理困境: 以英国教育罢工为例[J]. 深圳社会科学, 2023, 6(4): 139−149.
[38] 王森垚. 新自由主义的全球化困境及中国因应之道[J]. 理论探讨, 2021(5): 61−69.
[39] 丁晓钦, 罗智红. 美国新自由主义危机与当前经济“滞胀”风险: 以积累的社会结构理论为视角[J]. 教学与研究, 2022(9): 40−52.
[40] BIRHANE A. Algorithmic colonization of Africa[J]. Scripted, 2022, 17(2): 389−407.
[41] 戴克. 精英话语与种族歧视[M]. 齐月娜, 陈强, 译. 北京: 中国人民大学出版社, 2010.
[42] 李强. 当前我国社会分层结构变化的新趋势[J]. 江苏社会科学, 2004(6): 93−99.
[43] 姚何煜, 黄建. 社会学概论[M]. 成都: 电子科技大学出版社, 2019: 105.
[44] WEBER M. Class, Status, Party[C]// Grusky D B. Social stratification. Boulder: West view Press, 1994: 113−122.
[45] 李强. 社会分层十讲[M]. 北京: 社会科学文献出版社, 2011.
[46] CHUNG J. Racism in, racism out—A primer on algorithmic racism[EB/OL]. (2021−08−25) [2024−06−11]. https://www.citizen.org/article/algorithmic-racism/.
[47] 王传智. 数据要素及其生产的政治经济学分析[J]. 当代经济研究, 2022(11): 26−33.
[48] ENGLER A. Auditing employment algorithms for discrimination[EB/OL]. (2012−03−12) [2024−02−11]. https://www.brookings.edu/articles/auditing-employment-algorithms-for-discrimination/.
[49] 韦伯. 经济与社会: 上卷[M]. 林荣远, 译. 北京: 商务印书馆, 1997: 334.
[50] BOKÁNYI E, HANNÁK A. Understanding inequalities in ride-hailing services through simulations[J]. Scientific Reports, 2020, 10(1): 1−11.
[51] 甫玉龙, 刘杰, 鲁文静. 马克斯•韦伯社会分层理论视角下的美国贫困原因剖析[J].中国行政管理, 2015(4): 134−139.
[52] 许国亮. 政府权威研究[M]. 济南: 山东大学出版社, 2006: 232.
[53] 张强国. 法理型政党权威视角下政党协商发展问题探讨[J]. 广西社会科学, 2016(11): 145−149.
[54] 陈洪杰. 转型社会的司法功能建构: 从卡理斯玛权威到法理型权威[J]. 华东政法大学学报, 2023, 20(6): 57−71.
[55] 雷刚, 喻少如. 算法正当程序: 算法决策程序对正当程序的冲击与回应[J]. 电子政务, 2021(12): 17−32.
[56] BUOLAMWINI J. Artificial intelligence has a problem with gender and racial bias[EB/OL]. (2019−02−07) [2024−02−11]. https://time.com/5520558/artificial-intelligence-racial-gender-bias/.
[57] 王中原. 算法瞄准如何重塑西方选举: 算法时代的选举异化及其治理[J]. 探索与争鸣, 2021(5): 119−130, 179.
[58] STYPINSKA J. AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies[J]. AI & Society, 2023, 38(2): 665−671.
[59] ANDERSON M. Minority voices filtered out of Google natural language processing models[EB/OL]. (2022−12−09) [2024−02−11]. https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/.
[60] 傅春晖, 彭金定. 话语权力关系的社会学诠释[J]. 求索, 2007(5): 79−80.
[61] HARDING X. Breaking bias: Search engine discrimination? sounds about white[EB/OL]. (2021−09−28) [2024−02−11]. https://foundation.mozilla.org/en/blog/breaking-bias-search-engine-discrimination-sounds-about-white/.
[62] 姚何煜, 黄建. 社会学概论[M]. 成都: 电子科技大学出版社, 2019: 124.
[63] PARKIN F. Class inequality and political order: Social stratification in capitalist and communist societies[M]. London: Mac Gibbon & Kee, 1971: 50.
[64] 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[J]. 中外法学, 2017, 29(6): 1609−1623.
[65] 罗尔斯. 正义论[M]. 何怀宏, 译. 北京: 中国社会科学出版社, 2009.
[66] 弗里德曼. 人权文化: 一种历史和语境的研究[M]. 郭晓明, 译. 北京: 中国政法大学出版社, 2018: 139−140.
[67] 何怀宏. 公平的正义: 解读罗尔斯《正义论》[M]. 济南: 山东人民出版社, 2002.
[68] 奥维, 怀特. 欧洲人权法: 原则与判例[M]. 何志鹏, 孙璐, 译. 北京: 北京大学出版社, 2006: 478.
[69] 国务院新闻办公室. 2022 年美国侵犯人权报告[EB/OL]. (2023−03−28) [2024−02−13]. http://www.news.cn/world/2023-03/28/c_1129470457.htm.
[70] 杨. 正义与差异政治[M]. 李诚予, 刘靖子, 译. 北京: 中国政法大学出版社, 2017.

[71] Googlers Against Transphobia. Googlers Against Transphobia and Hate[EB/OL]. (2019−04−01) [2024−02−13]. https://medium.com/@against.transphobia/googlers-against-transphobia-and-hate-b1b0a5dbf76.

[72] 美国侵犯难移民人权的事实真相[EB/OL]. (2023−03−30) [2024−02−13]. https://www.mfa.gov.cn/web/wjbxw_new/202303/t20230330_11051564.shtml.

[73] WILLIAMS B A. Tech's troubling new trend: Diversity is in your head[EB/OL]. (2017−10−19) [2024−02−13]. https://www.nytimes.com/2017/10/16/opinion/diversity-tech-women-silicon-valley.html.

[74] BENSON A, LI D, SHUE K. Potential and the gender promotions gap[EB/OL]. (2023−07−24) [2024−02−13]. https://doi.org/10.5465/AMPROC.2023.19580abstract.

[75] McKinsey & Company. Women in the workplace 2023[EB/OL]. (2023−10−05) [2024−02−14]. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace-2023.

[76] 沈岿. 论软法的实施机制: 以人工智能伦理规范为例[J]. 财经法学, 2024(6): 108−127.

[77] 曾雄, 梁正, 张辉. 欧盟人工智能的规制路径及其对我国的启示: 以《人工智能法案》为分析对象[J]. 电子政务, 2022(9): 63−72.

[78] 马长山. 智能互联网时代的法律变革[J]. 法学研究, 2018, 40(4): 20−38.

Towards algorithmic justice: The social construction of algorithmic discrimination and its governance strategies

MAO Junxiang[1], GUO Min[2]

(1. School of Law, Central South University, Changsha 410083, China;
2. School of Humanities, Central South University, Changsha 410083, China)

Abstract: Algorithm, as a socially embedded entity, interacts dynamically with the social structure. Algorithmic discrimination emerges as a socially constructed phenomenon shaped by the historical legacy of societal discrimination, the structural barriers to algorithmic correction, the subtle infiltration of value biases, and the interest-driven orientation of the social ecosystem. This discrimination does not represent a novel form of inequality inherent to algorithmic technology itself, but rather a reflection of enduring historical and contemporary issues in the algorithmic age. Algorithmic discrimination impedes equitable social mobility across dimensions of wealth, power, and prestige, thereby exacerbating imbalances within social structures. Addressing this necessitates constructing "algorithmic justice" to counter algorithmic discrimination. Establishing algorithmic justice requires a dual commitment to distributive and relational justice, integrating protected characteristics into algorithmic decisions, respecting group diversity without weaponizing it, acknowledging the need for special accommodations based on group disparities, and amplifying the decision-making voice of minority groups in algorithmic contexts. Realizing algorithmic justice should adopt a tripartite governance model which encompasses technology, law, and ethics, with a focus on designing legal frameworks centered on algorithmic distinctions and implementing a comprehensive AI governance mechanism grounded in "technology for good."

Key words: algorithmic discrimination; algorithmic justice; algorithmic distinctions; algorithmic governance

[编辑：苏慧]