交大法学 SJTU Law Review
No. 1 (2024)

算法歧视的两副面孔及其法律规制

李志颖*

目次

一、现行算法歧视法律治理的两个弊端
二、弊端的消除：厘清算法歧视的两副面孔
三、“算法偏见”的法律规制
四、“算法不平等”的法律规制
五、结论

摘要 我国现行算法歧视的法律治理存在两个弊端，分别是算法歧视规制范围过宽和算法歧视规制工具的凌乱。弊端的消除应当追溯至歧视概念本身，通过区分歧视的法律维度和道德维度，呈现算法歧视的两副面孔。第一副面孔是“算法偏见”，在语义上采用法律维度的歧视，对其治理旨在解决算法当中的历史偏见问题，适用反歧视法的规制原理，对算法设计者和算法应用者施加基于社会法的结果义务；第二副面孔是“算法不平等”，在语义上采用道德维度的歧视，对其治理旨在解决信息时代的资源分配问题。对于“算法不平等”，既要尊重私法自治又要削弱潜在歧视风险，法律规制除了事后司法救济外还可以采取事中方案，通过价值维度和信任维度，对算法设计者和算法应用者施以适度的负担。

关键词 算法歧视 算法偏见 历史偏见 算法不平等 资源分配

在信息时代混沌初开之际，“算法歧视”作为一个算法领域的主要问题被提出，为国际组织和各国学者所承认和关注，并被讨论至今。算法歧视系指在大数据、人工智能和自动化决策等技术语境下，机器所做出的决策对个体或群体产生不合理的区别对待。算法歧视现象广泛存在且各式各样，既有在雇佣、教育、医疗健康、量刑和警力分配等场景对弱势群体的人格歧视，也有遍布日常生活的信用评分、个性化定价和推荐等经济型区别对待。法律如何规制算法歧视顺势成为研究热点。法律规制的目的除了填补实际损害外，还在于削弱潜在风险，为此规制覆盖算法决策的设计、应用等全过程。

然而，我国既有的对算法歧视的法律研究或是忽略算法歧视现象的多样性，或是忽略不同算法歧视现象的内在关联，使得算法歧视的法律规制存在弊端。既有研究的规制策略主要分为两类：一是“一把抓”式地规制算法歧视。〔1〕这类研究即便对算法歧视的分类有所涉及，也仍是笼统

* 复旦大学法学院博士研究生。

〔1〕 参见崔靖梓：《算法歧视挑战下平等权保护的危机与应对》，载《法律科学》2019 年第 3 期；张莉莉、朱子升：《算法歧视的法律规制：动因、路径和制度完善》，载《科技与法律》2021 年第 2 期。

地提出算法影响评估、算法透明、算法解释、算法正当程序等规制方案，不加区分地适用于所有算法歧视现象。这类成果往往属于早期研究，为算法歧视的法律研究开疆拓土。二是沦为“具体情况具体分析”的对策性规制。〔2〕随着数字法学的精细化研究，许多学者转向特定现象的算法歧视，虽然一定程度上摆脱了早期研究的泛化和笼统，但却使得算法歧视的法律规制成为对策性的而非规范性的，漠视了不同算法歧视之间的内在关联。

这将导致现行算法歧视法律治理存在两个弊端，即算法歧视规制范围过宽和算法歧视规制工具的凌乱。该问题的解答应当追溯至歧视概念本身，在此基础上本文区分了算法歧视的两副面孔，分别是“算法偏见”和“算法不平等”。根据对应不同的规制原理，具体回答规制的方案、时机和规制程度，有助于整体性地处理算法歧视问题。

一、现行算法歧视法律治理的两个弊端

对于算法歧视，现行法律治理存在两个弊端：一是治理对象的泛化，即算法歧视规制范围过宽；二是治理手段的碎片化，即算法歧视规制工具的凌乱。

（一）治理对象的泛化：算法歧视规制范围过宽

由于概念的模糊，歧视往往与平等、公平和正义相关联，导致规制者对算法歧视作泛化对待。〔3〕具体表现如下：一是算法决策只要存在区别对待就被视作算法歧视。这种做法无异于以正义之名实施过度干预，比方说，“大数据杀熟”不论合理与否一律被贴上算法歧视的标签；二是把一切的数字不公视作算法歧视，此时算法歧视已不局限于算法决策所产生的不合理区别对待，还包括基于算法权力产生的数字鸿沟、数据垄断、平台异化、数据茧房等数字不平等。〔4〕规制者应当根据算法性质、处理对象、场景的不同适用相对应的规制原理和工具，不合理地扩大规制范围将会导致规制原理和工具的错乱。以算法风控系统名誉侵权案为例，原告通过涉案软件与他人聊天时因频繁出现“私募”“基金”等内容，致使其账号被判定为高风险并封禁。〔5〕有学者将其置于算法歧视的讨论当中，〔6〕本文认为这种处理方法泛化了算法歧视的规制范围，导致不合适的规制工具被纳入算法歧视的法律规制之中。

算法歧视是一项具体的风险，治理算法歧视只是实现算法公平的其中一环。为此，应当围绕算法歧视建立具体制度，明确其规制范围而非作泛化处理。美国白宫科技政策办公室发布的《人工智能权利法案》所主张的“算法歧视保护”原则（Algorithmic Discrimination Protections）便抛弃了泛化处理，只保护基于种族、肤色、民族、性别、宗教等法定特征的不合理区别对待和影响。〔7〕

〔2〕 譬如，专门研究传统人格权意义的算法歧视，参见李成：《人工智能歧视的法律治理》，载《中国法学》2021年第2期；专门研究女性就业的算法歧视，参见阎天：《女性就业中的算法歧视：缘起、挑战与应对》，载《妇女研究论丛》2021年第5期；专门研究大数据杀熟的，参见王潺：《“大数据杀熟”该如何规制？——以新制度经济学和博弈论为视角的分析》，载《暨南学报（哲学社会科学版）》2021年第6期。

〔3〕 参见宁园：《算法歧视的认定标准》，载《武汉大学学报（哲学社会科学版）》2022年第6期，第154—155页。

〔4〕 参见张吉豫：《数字法理的基础概念与命题》，载《法制与社会发展》2022年第5期，第52页。

〔5〕 参见《2022年度中国十大传媒法事例简介及入选理由》，载中国法院网2023年1月7日，https://www.chinacourt.org/article/detail/2023/01/id/7091844.shtml。

〔6〕 参见袁文全：《算法歧视的侵权责任治理》，载《兰州大学学报（社会科学版）》2023年第2期，第95—96页。

〔7〕 See *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, the White House, October, 2022, p.10.

我国也应当结合本土实际情况，明确算法歧视的规制范围。

（二）治理手段的碎片化：算法歧视规制工具的凌乱

为了治理算法歧视，除了司法救济作为兜底工具之外，还有算法影响评估、算法解释、算法正当程序、反对自动化决策等规制工具被提出。现行算法治理规范几乎没有专门针对算法歧视的法律条款，在《互联网信息服务算法推荐管理规定》草案的用户标签条款和《个人信息保护法》二审稿的敏感个人信息条款中一度存有针对“歧视”的规定，但最终在现行法中被剔除。为此，算法歧视的治理只能依附于相关算法治理法律规范的既有制度，典型如算法解释、算法影响评估和主张各类信息权利等。同样地，部分学者提出的治理方案，不论是治理何种类型的算法歧视，仍然是对这些规制工具的反复重申。〔8〕这些规制工具俨然已成为信息时代的“大道理”和“万灵药”。即便如此，我们也不禁要问，面对广泛存在且各式各样的算法歧视，这些规制工具是否每一项都卓有成效，以及是否每一项都要适用，还是有针对性地适用？适用的程度是否一致？如果根据算法歧视的不同类型适用不同方案的规制，那么规制的背后是否存在体系性的思考？现有规制往往是笼统地讨论算法歧视或以具体问题具体分析的方式为特定算法歧视给出对策，难以有效回答上述问题，致使算法歧视治理手段碎片化。

治理手段的碎片化将导致算法歧视风险分配的不合理，并给算法设计者和算法应用者带来沉重的负担。一是徒劳浪费算法设计者和算法应用者的规制精力。上述“大道理”规制工具并非放之四海而皆准，有的规制工具在特定领域收效甚微。算法解释便是如此，只适合治理特定领域的算法歧视，下文将详细说明。二是付出高昂的合规成本。不同的规制工具不仅在削减算法歧视的能力上不同，在规制成本上亦有所差异。以算法影响评估为例，该方案往往被视作有力的规制工具，并以《个人信息保护法》第56条为其背书。然而，该方案的成本相当高昂，根据调查研究，每一个算法应用的年度评估费用可高达29277欧元。〔9〕这些合规成本必然会使中小型企业承担不可承受之重。为此，算法影响评估应当置于恰当的领域，如欧盟《一般数据保护条例》、美国《算法问责法案》的算法影响评估主要聚焦于公共决策和高风险领域。三是牺牲算法的应有价值。算法歧视的规制成本必然转移至算法设计者和算法应用者，过重的规制成本将严重降低算法的效率。除此之外，削减算法歧视与实现算法精准性往往处于此消彼长的平衡关系。〔10〕为此，不恰当的治理手段往往以效率、计算精准性为代价，不但可能使得算法失去其应有价值，同时可能大大抑制科技创新。

二、弊端的消除：厘清算法歧视的两副面孔

导致上述治理弊端的主要原因在于规制者没有清晰认知歧视，不知何为歧视，何以讨论算法歧视及其法律规制。在治理对象上，歧视概念本身模糊不清，致使算法歧视规制范围泛化；在治理手段上，上述规制工具与歧视概念本身是脱节的，仅仅是把“算法的法律规制”议题中老生常谈的方案挪至算法歧视法律规制之中。同时，不加区分地规制算法歧视也反映了不同类型的歧视被混

〔8〕 参见石颖：《算法歧视的缘起、挑战与法律应对》，载《甘肃政法大学学报》2022年第3期，第64—68页。

〔9〕 See *Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe*, European Commission, 2021, p.134.

〔10〕 See Pak-Hang Wong, *Democratizing Algorithmic Fairness*, 33 Philosophy & Technology 225, 229 (2020).

为一谈。为此，算法歧视的探讨应当首先追溯至歧视概念本身，通过区分歧视的法律维度和道德维度，以厘清算法歧视的两副面孔，本文将这两副面孔称为“算法偏见”和“算法不平等”。只有在明确治理对象之后，才能根据相应的规制原理采取恰当的治理手段。

（一）辨析与区分：法律维度的歧视与道德维度的歧视

“歧视”是一个容易被误用的语义词汇，在语义表达上存在道德维度与法律维度的区分。〔11〕道德哲学家往往把“歧视”定义为“不合理区别对待”，基于任何特征的不合理区别对待都会被视作歧视。〔12〕道德维度的歧视在理解上趋近于不平等，该理解也常为日常生活所用，这也是为什么歧视与不平等往往并列使用。但是，在法律维度里，歧视是一个具体的法律概念，只是法律不平等的子集之一。“歧视”的法律通行定义为基于种族、性别、年龄、信仰、残疾等特征做出的区别、排斥、优惠或限制。通过上述两个定义的比较可以发现，道德维度的歧视与法律维度的歧视是不尽相同的。然而，我们仍然难以区分这两个维度，主要原因在于“歧视”的语义表达往往在不同法律场景中被混淆使用，造就关于歧视的“语义学之刺”。

“歧视”的语义表达及其法律调整具体如图 1 所示。在日常生活中，我们所谈的“歧视”在语义表达上属于道德维度的歧视，仅仅强调不合理的区别对待。不合理的区别对待理所当然为道德所唾弃，但并非必然受到法律约束，只有部分不合理的区别对待由法律调整。当法律要调整不合理区别对待的时候，往往是通过领域法的方式展开，以“特殊——一般”的关系列举如下：通过反歧视法来调整，对应法律词典和国际条约关于歧视的理解，在语义上对应的是法律维度的歧视。反歧视法是为了确保个体摆脱历史偏见的约束而顺利融入社会，使得个体能够“正常”地生活在社会当中。〔13〕基于其重要性，反歧视法作为歧视现象的特别领域法而被单独列出。除此之外，还有各式各样的不合理区别对待通过其他领域法来调整，如竞争法上的歧视、价格法上的歧视等等。但对于这些歧视，人们在语义表达上仍是采取道德维度的歧视，笼统地套上不平等的高帽。综上，人们日常惯用的“歧视”语义表达存在不理性的情况，或把所有歧视问题置于法律范畴来谈，或把不同

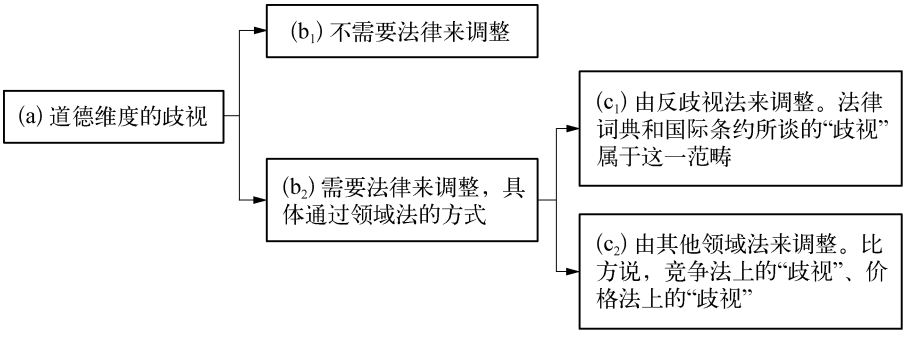

图 1 “歧视”的语义表达及其法律调整

〔11〕参见李志颖：《论歧视的法律定义——基于社会行为视角的分析》，载《法制与社会发展》2023 年第 1 期，第 129—130 页。另外，还存在类似的观点，即通过是否以法律专业人的视角来审视歧视，区分了法律层面的歧视（discrimination in law）和朴素层面的歧视（discrimination in lay）。See Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2015, p.1－4.

〔12〕See Kasper Lippert-Rasmussen, *Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination*, Oxford University Press, 2014, p.15; Benjamin Eidelson, *Discrimination and Disrespect*, Oxford University Press, 2015, p.17.

〔13〕见前注〔11〕，李志颖文，第 129—130 页。

领域的歧视问题混为一谈，或“歧视”语义表达与落脚领域不一致。本文的首要工作是揭示被不理性表达所掩盖的应然秩序。

对于算法歧视，“歧视”一词正是在“作为特定的法律不平等之一”（在语义上采用法律维度的歧视）和“宽泛意义的不平等”（在语义上采用道德维度的歧视）之中被混淆使用。区分法律维度与道德维度的歧视，“算法歧视”一词呈现两副面孔，即“算法偏见”和“算法不平等”。

（二）对应法律维度歧视的“算法偏见”

算法歧视的第一副面孔为“算法偏见”，在语义表达上对应法律维度的歧视，在关系图上落脚在c_1，它所讨论的是历史偏见在算法场景的运用。正如《中国关于加强人工智能伦理治理的立场文件》指出，人工智能技术的误用滥用将会加剧“歧视和偏见”。“算法偏见”的典型场景如：在卢米斯案（State *v*. Loomis），法官运用COMPAS算法软件进行刑事量刑，COMPAS以性别和种族作为评价指标的量刑预测引发了算法歧视的法律热议；亚马逊公司所使用的HireVue招聘算法存在性别歧视，如果求职者毕业于女校或简历带有女性字样，将会受到算法的负面评价。仔细品读这些典型场景，算法歧视可能潜藏于数据的收集、筛选和训练当中，也可能出现在算法的设计和运行过程当中，但深层原因在于历史偏见在民生与公共服务等重要领域作祟。历史偏见有两个特征：一是从个体角度而言，历史偏见是个人无意识的社会认知。内群起初对外群的恶意揣度和偏见将随着历史的潜移默化成为社会认知的前理解。二是从社会角度而言，历史偏见演化为社会结构性不平等。历史偏见影响不同人群间的分配格局，调整资源配置以及重塑社会权力结构，以独立的运作方式成为社会结构的一部分。〔14〕正是历史偏见的显著与根深蒂固，个体或群体受制于性别、年龄、种族等“第一印象”而无法融入与民生事项和公共服务等相关的重要社会交往，如教育选择、工作就业、社会福利、政治参与等。历史偏见作为本源，借由大数据和人工智能等科技载体的粉饰而摇身一变，形成“算法偏见”。即便如此，“算法偏见”看似新颖，其讨论逻辑仍置于法律维度的歧视，即反歧视法框架之内。

（三）对应道德维度歧视的“算法不平等”

算法歧视的第二副面孔为“算法不平等”，在语义表达上对应道德维度的歧视，但这一日常语义表达往往是不严谨的，在关系图上实际落脚在c_2，它所讨论的是算法场景中的资源分配。“算法歧视”与“算法不平等”两个概念经常被并列使用：2018年5月多个国际组织联合发表关于智能算法规制的《多伦多宣言：机器学习系统中平等和非歧视权利的保护》，该宣言标题的用词为“平等和非歧视”；2019年4月欧盟发表《可信赖人工智能的道德准则》明确了七个核心要素，其中之一为“多元化，非歧视和公平”〔15〕。这时候，算法歧视在语义上采用的是道德维度而非法律维度的歧视，展现为“算法不平等”，其所讨论的内容不局限于历史偏见，而是结合资源分配来讨论更为宏大的公平议题。下述这些议题与“算法不平等”现象紧密结合：数据能力的不平等议题，或称作“数字鸿沟”。〔16〕除此之外，还有分配正义的议题，〔17〕充当“立法者”的算法能否做到同等情况同等对待、不同情况差别对待，背后所涉及的则是“机会平等”“利益平衡”，这些词汇在算法歧视的讨论中

〔14〕 见前注〔2〕，李成文，第130页。

〔15〕 *High-level Expert Group on Artificial Intelligence*, *Ethics Guidelines for Trustworthy AI*, European Commission, 8 April, 2019, p.18－19.

〔16〕 参见陈林林、严书元：《论个人信息保护立法中的平等原则》，载《华东政法大学学报》2021年第5期，第10—11页。

〔17〕 参见高富平：《个人信息使用的合法性基础——数据上利益分析视角》，载《比较法研究》2019年第2期；程金华：《利益平衡：“三位一体”的个人信息保护法律治理架构》，载《探索与争鸣》2020年第11期。

频频出现。现阶段,“算法不平等”现象主要集中在商业领域的自动化决策,如“大数据杀熟”下用户黏度更高的人群需要支付更贵的价格,以及“数字信贷歧视”下信贷机构通过设置不合理的特征输入使得个别群体被画像为不适格者而无法借贷。这些基于特征的筛选结果和资源分配虽然都被视作不合理区别对待,但与“算法偏见”大不相同的是,这里的特征与历史偏见无关,而运用的场景私法色彩更为浓厚。

(四)小结:“算法偏见”与“算法不平等”的区别

算法歧视的触发机制遍布算法全过程:在输入环节中,收集的数据存在偏见以及代表性不足,算法设计者无意识嵌入歧视目的使得算法变成歧视性产品,更为致命的是代理特征(proxy)悄无声息携带歧视信息;在学习环节,机器挖掘和深度学习不可避免复制上述环节的错误信息,而算法设计者也无法清晰了解算法每一步是如何运作的,使得在输出环节“偏见进、偏见出”。〔18〕但是,这么一套机制既可以产生“算法偏见”,也可以产生“算法不平等”。唯有厘清两者之间的区别,才能对症下药。

相较于传统的算法场景划分,基于“法律维度的歧视”与“道德维度的歧视”两类语义表达的场景划分之最大优势在于符合理论的周延性和实践的多样性。对于场景划分,主流基调是划分为公法和私法,〔19〕除此还有实用主义的划分,〔20〕即“算法进入公权力”与“算法进入商业领域”。无论是主流基调还是实用主义的划分仍然坚持公私划分的绝对壁垒,然而算法歧视的现象列举早已对其提出挑战,譬如自动化雇佣虽属于私领域,但应当划入类公法范畴。为此,区分歧视的法律维度和道德维度,算法歧视的场景划分既能符合理论的周延性,又能融贯实践的多样性。如果要用公私法标准来描述的话,“算法偏见”的运用场景主要是公法领域,也包括部分关切民生与个人福利的私法领域;“算法不平等”的运用场景则是其余私法领域。

综上,“算法偏见”与“算法不平等”作为算法歧视的两大类型,具有以下区别:(1)解决对象和适用范围不同:前者的解决对象是历史偏见,仅作用在有限特征及与民生和自我实现息息相关的重要社会交往领域;后者的解决对象是资源分配,不局限于任何特征,主要适用于部分私领域。(2)性质不同:前者定性为“社会法”范畴,秉持特定的实质平等;后者在性质上关涉的是“反歧视原则进入私法领域”。例如,“大数据杀熟”仅是普通的私法定价问题,由于它违反反歧视原则而引起争议,但放入“算法偏见”来讨论则格格不入。(3)对应的规制原理及法律制度不同:前者既然对应作为法律维度的歧视,理所应当运用反歧视法来规制;后者的讨论需要细化到具体场景并由与之相关的领域法来填充价值维度,譬如“大数据杀熟”由反垄断法来规制。

三、“算法偏见”的法律规制

“算法偏见”在语义上对应法律维度的歧视,其法律规制理所应当采取反歧视法,以解决历史偏见事宜。纵观各国反歧视立法与司法实践,反歧视法的基础理论区分强调意图的直接歧视和强调效果(亦称为结果)的间接歧视。对于传统歧视行为而言,法律主要采用直接歧视理论,对少部

〔18〕 参见岳平、苗越:《社会治理:人工智能时代算法偏见的问题与规制》,载《上海大学学报(社会科学版)》,2021年第6期,第2—3页;Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 California Law Review 671, 677-693 (2016).

〔19〕 参见林洹民:《〈个人信息保护法〉中的算法解释权:兼顾公私场景的区分规范策略》,载《法治研究》2022年第5期。

〔20〕 参见张凌寒:《商业自动化决策的算法解释权研究》,载《法律科学》2018年第3期,第66页;张欣:《算法解释权与算法治理路径研究》,载《中外法学》2019年第6期,第1441页。

分歧视行为采用强保护的间接歧视理论。在信息时代，新的外部环境促使反歧视法做出理论革新，算法歧视的规制原理应当采用间接歧视理论，而非直接歧视理论。不少算法歧视的研究就是在这个逻辑下展开的。〔21〕

（一）直接歧视的失败：无法识别“算法偏见”的歧视意图

直接歧视的核心在于歧视意图，即通过发现存在歧视意图来证明发生了歧视行为。以卢米斯案为标志，该案印证了“算法偏见”中的歧视意图是难以被发现的，直接歧视的规制路径面临挑战。〔22〕下文结合卢米斯案，来说明在“算法偏见”之中为什么难以发现歧视意图。

首先，歧视意图藏匿于算法黑箱之中。算法本身是没有意志的，只是根据既定的编程来演算和输出结果。算法虽然作为中立工具却承载着意识，意识或来自算法设计者，或来自数据。〔23〕但是，算法黑箱遮蔽了算法所承载的意识，使得人们难以辨识这些意识的具体内容及其出处。对于算法决策来说，算法黑箱是必然存在的。根据透明性和可解释性的程度，算法黑箱可区别为强版本和弱版本。〔24〕对于强版本的算法黑箱来说，人们只知道输入端和输出端，对于如何计算一无所知，自然而然也不可能识别歧视意图。弱版本的算法黑箱是对算法可解释性和透明性的让步，一定程度揭露了算法的考量指标及其权衡比重序列，但考虑到商业秘密保护不可能完全披露计算公式。即便如此，弱版本的理解所提供的计算信息也只是模糊的，难以准确判断输出结果和计算指标的因果关系，难以识别是否存在歧视意图。在卢米斯案中，卢米斯意识到性别是计算指标，故主张COMPAS是性别歧视的。但是，法院认为，性别的使用只是为了增加计算的精准性，仅仅依据“性别作为计算指标”不足以认定COMPAS具有性别歧视意图。〔25〕

其次，代理特征隐藏歧视意图。在卢米斯案中，法官为了讨论COMPAS是否产生性别歧视，引用了科研机构对COMPAS种族歧视的实验研究，承认该算法产生了种族方面的不利歧视影响，使得黑人群体被认定为具有更高的犯罪风险。〔26〕即便COMPAS在结果上是种族歧视，我们回过头审视COMPAS的算法指标，仍然无法判断其存在歧视意图。COMPAS算法虽然没有直接使用“种族”作为计算参数，但使用了“教育”“职业”“犯罪记录”等代理特征，间接反映了个体的种族信息。换句话说，即便在输入端完全剔除种族、性别、肤色等受保护特征，大数据仍然能够借助其他替代性数据，隐藏歧视意图，实现歧视结果。

最后，歧视意图被其他目的或价值所掩盖。在科学世界里，算法的计算精准性可谓在至高位阶，致力于如实地孪生和预测现实。正如上文所述，性别作为COMPAS计算参数，是服务于算法的精准而非歧视目的。另外，切入具体的人文运用场景，歧视意图或许会被算法所追求的实用目的所掩盖，如为了国家利益和社会秩序，大部分国际机场通过算法筛选并对特定人群进行更严格

〔21〕 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》，载《比较法研究》2019年第4期；张恩典：《反算法歧视：理论反思与制度建构》，载《华中科技大学学报（社会科学版）》2020年第5期；Solon Barocas & Andrew D. Selbst, *supra* note〔18〕, at 671 - 732; Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 Alabama Law Review 519, 519 - 572(2018).

〔22〕 See *Beyond Intent: Establishing Discriminatory Purpose in Algorithmic Risk Assessment*, 134 Harvard Law Review 1760, 1768 (2021).

〔23〕 参见伊卫风：《算法自动决策中的人为歧视及规制》，载《南大法学》2021年第3期，第81—88页。

〔24〕 See Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harvard Journal of Law & Technology 889, 905 - 906 (2018).

〔25〕 See State *v*. Loomis, 881 N. W. 2d 749 (Wis. 2016), at 766 - 767.

〔26〕 Ibid., at 764.

烦琐的检查；[27]或被其他价值所掩盖，譬如算法商业必要性，算法本身容易成为借口来合理化雇主的歧视行为，把歧视行为高度伪装成与工作所需密切关联，以成立商业必要性来规避歧视责任。[28] 当然，歧视意图还会被商业秘密所掩盖。

（二）通过间接歧视规制“算法偏见”：施加结果义务

既然难以识别“算法偏见”其中的歧视意识，只能另寻出路。以在结果上产生客观歧视效果并达到差别性影响作为歧视认定标准的间接歧视则大有用武之地，间接歧视可谓解决产生歧视效果但“表面中立”的同一对待行为的灵丹妙药。即便如此，也不乏反对的声音认为间接歧视对“算法偏见”的规制存在缺陷。其中一个看似致命的反对观点为：间接歧视只关注群体公正而不利于调整个性化的算法决策。[29] 但是，如果明晰法律维度的歧视与道德维度的歧视之区分，承认“算法偏见”的解决对象主要是历史偏见，那么这种反对观点便不成立，相反个性化算法决策的公平问题应当放到“算法不平等”里头。除此之外，间接歧视还可能存在使用泛滥等问题，但这些缺陷都不致命，只需理论修正即可，学界普遍认为间接歧视是更适合解决“算法偏见”的理论框架。[30] 为此，当我们认同结果论的间接歧视作为解决之道，沿着间接歧视的逻辑要求，算法设计者和算法应用者应当承担某种结果论的义务，保证在结果上不会产生歧视影响。这也是对“算法偏见”施以事前规制的合法性依据。当算法设计者和算法应用者履行结果义务后仍然产生歧视问题，则诉诸司法救济。其司法救济仍然以间接歧视为主要框架，主要讨论损害的认定、过错的认定和举证责任分配等如何向被歧视者倾斜。不少成果已对“算法偏见”的司法救济展开了深入研究，[31]下文遂主要探讨为了削弱歧视风险，可以施以何种以及多大的结果义务，以及为什么应当施以结果义务。

算法设计者和算法应用者的结果义务往往着手于算法设计阶段或运用之前的评估阶段，属于事前规制，主要可归为两类。一类是算法影响评估，即评估算法决策产生的社会效果。正如上文所言，算法影响评估是高成本的规制手段并主要适用在公共决策领域。基于成本考量，部分“算法偏见”如企业的自动化雇佣并不适用算法影响评估，需要另寻方案。另一类方案本文称之为“数据控制”，即对受保护特征及其代理变量进行特殊处理，通过参数的控制达成结果的控制。在法律世界里，这种数据控制是对输入端的数据控制，而不应该是输出端。科学家往往在输出端下功夫，把结果义务转化为通过设计关于“非歧视”“公平”的数理公式，以确保输出端上的结果平等。举例说明，科学家如果根据“人口统计平等”（demographic parity）来设计数字贷款算法，那么算法在输出结果上应当保证男性与女性获得贷款的概率相等；如果根据“机会平等”（equality of opportunity）来设计数字贷款算法，那么算法在输出结果上应当保证男性获得贷款的预测准确率与女性的预测准确率相等。这里的问题在于，何种结果才是平等以及结果平等是否真的平等？法律人思维与科学家思维是有所差异的。反歧视法从不要求输出端的结果平等，而是关切输入端涉嫌歧视的争议信息。[32] 无论何种歧视案件，对于争议信息只要可以论证其充分性和合理性，即可

〔27〕 参见沈伟伟：《算法透明原则的迷思——算法规制理论的批判》，载《环球法律评论》2019 年第 6 期，第 25—26 页。

〔28〕 See Solon Barocas & Andrew D. Selbst, *supra* note〔18〕, at 709.

〔29〕 见前注〔21〕，张恩典文，第 65 页。

〔30〕 See Stephanie Bornstein, *supra* note〔21〕, at 542 - 543.

〔31〕 See Solon Barocas & Andrew D. Selbst, *supra* note〔18〕, at 671 - 732; Stephanie Bornstein, *supra* note〔21〕, at 519 - 572.

〔32〕 See Robert Bartlett, Adair Morse, Nancy Wallace & Richard Stanton, *Algorithmic Discrimination and Input Accountability under the Civil Rights Acts*, 36 Berkeley Technology Law Journal 675, 685 (2021).

不被视作歧视。输出端的歧视结果只会在程序上影响被告承担更重的论证责任，而不会直接以结果成立歧视。为此，法律人往往会关注输入端的数据控制。算法歧视法律研究的初期，最为常见的输入端控制方案为对受保护特征的去盲或完全剔除，但这种方案严重影响算法计算的准确性。从科学的角度而言，我们不应当排除而是积极使用敏感信息，以确保算法决策的科学性。在这个意义上，《互联网信息服务算法推荐管理规定》的删除反映个人特征的用户标签和《个人信息保护法》禁止使用敏感信息等相关规定并不适合规制“算法偏见”。鉴于此，部分学者提出不排斥敏感信息的控制方案，如“输入义务测试”〔33〕“二阶段去偏见法”〔34〕。数据控制方案往往都带有较强的传统反歧视法身影，而非完全科学计算的结果。法律究竟采用何种数据控制方案，是对科学性、可操作性、信息处理者成本等因素综合考量的结果。

那么，这种结果义务源于何处？为什么要施加在算法设计者和算法应用者身上呢？其正当性主要有两点：第一，这是一种责任转移(burden-shifting)的义务。〔35〕反歧视理论的直接歧视与间接歧视，实则借鉴于侵权责任。根据侵权责任原理，原告首先应当举证，证明某算法存在受保护特征的明显分类并带有歧视意图。但是，“算法偏见”的特性使得原告无法识别其中的歧视意图，把原告陷于严重不公地位。这时候将举证责任倒置，由算法设计者和算法应用者来证明算法不存在歧视影响。〔36〕第二，更为核心的是，这是社会法上的义务。社会法导入“身份”因素，以基准法的形式为弱势群体提供特殊保护，对相对强势一方课以法定义务，以校正契约自由在两者之间造成的偏差。〔37〕反歧视法作为社会法，通过施加强制义务以保证弱势群体能融入与民生和自我实现密切相关的社会交往。这一要求同样适用于由算法掌控的社会交往，算法设计者和算法应用者承担相应的社会法义务。当然，也有学者提及对弱势群体予以倾斜保护这一理念，譬如算法歧视可以依赖侵权责任的过错推定，以及合同责任的先合同义务和附随义务，实施民法救济，其中过错推定、先合同义务以及附随义务，实质上都是合同责任和侵权责任对弱势一方的倾斜保护。〔38〕但由于这里需要解决的是历史偏见，理所应当采取强倾斜的社会法保护，对算法设计者和算法应用者施以强规制；相反，民法救济这种弱倾斜保护，更适用于解决资源分配方面的“算法不平等”。

对于义务的来源和正当性，也许会有学者提出第三种声音，即义务来自算法透明和可解释性的要求。在信息时代，不可否认算法透明和算法解释占据十分重要的地位。然而，对于这种声音，本文并不认同。算法透明和可解释性的实现虽然在理论上有助于规制“算法偏见”，但事实上难以

〔33〕 该方案大致内容为：首先确定为实现算法目的而必要的目标变量，并以此作为标准。然后将争议变量的计算结果分解为两部分：一部分是可靠部分，因为这部分发挥着与目标变量相同的作用；另一部分为错误部分。这时候，就要评估错误部分是否与歧视信息相关；如果与歧视相关，则不通过输入义务测试，该争议变量为歧视信息。Ibid., at 681.

〔34〕 该方案大致内容为：在评估阶段消除代理特征带来的歧视影响，经过第一步的评估阶段，那么在预测阶段输入端已经没有个体层面的歧视信息。See Crystal S. Yang & Will Dobbie, *Equal Protection under Algorithms: A New Statistical and Legal Framework*, 119 Michigan Law Review 291, 297 (2020).

〔35〕 See Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 Georgia Law Review 109, 161, 193 (2017).

〔36〕 类似于《个人信息保护法》第69条，个人信息处理者承担过错推定，其原因在于个人信息处理者与自然人的举证能力之差异。参见程啸：《侵害个人信息权益的侵权责任》，载《中国法律评论》2021年第5期，第63页。

〔37〕 参见余少祥：《社会基准法的本质、特征及其实施机制》，载《北京大学学报(哲学社会科学版)》2019年第3期，第138页。

〔38〕 参见潘芳芳：《算法歧视的民事责任形态》，载《华东政法大学学报》2021年第5期。

充当“算法偏见”规制义务的核心。换句话说,“算法偏见”的规制不依赖于算法透明和可解释性。落实算法透明与可解释性,意味着阐明因果关系。无论是采用何种立场和理论的因果关系,在实际运转层面,算法黑箱都是难以跨越的障碍。〔39〕即便技术攻克了算法黑箱,使算法得以被规制,这种技术也成为“屠龙刀”,使算法的价值大打折扣。〔40〕相反,本文认为,因果关系并非必须攻克的课题。因为社会法的强制义务是对弱势群体的特别倾斜,以至于我们可以弱化因果关系。因果关系的弱化在社会法中并不罕见。譬如美国反歧视司法实践采用“五分之四原则”来判断是否达到歧视程度。具体来说,在雇佣领域,如果受保护群体的录取率低于其他群体录取率的五分之四,则可以认为此雇佣标准具有不利的后果,原告据此可以形成初步证据。〔41〕即便五分之四原则早已被统计学所证伪,但不妨碍司法实践继续适用。同样地,根据我国《劳动合同法》的规定,未与员工签订劳动合同,雇主将按月支付双倍工资。很显然,这里的赔偿标准不是按照成本收益的数理计算,而是强调对弱势群体的特殊保护。正是如此,“算法偏见”的法律规制可以采取因果关系的弱化方案,通过人文制度方案而非技术方案来实现歧视结果的控制。但是,“算法不平等”也会对算法设计者和算法应用者施加义务要求,但这个义务则依赖于因果关系。

四、“算法不平等”的法律规制

(一)“算法不平等”的规制时机

“算法不平等”这种算法歧视在语义上采用道德维度的歧视,其表达的背后充斥着不理性因素。“算法不平等”现象是否真的不平等,往往是有争议的。以数字信贷为例,借贷作为经济行为理所应当追求收益最大化,并结合安全性、流动性和效益性选择“优质客户”。使用性别、年龄、种族等传统歧视特征作为筛选标准,必然为道德所唾弃,但我国法律是否禁止并无定论。在信息时代,信贷机构利用算法广泛采集移动货币账户的存款、转账、商户支付和账单支付活动、社交媒体、短信和互联网浏览活动等行为数据和身份数据,来分析借款人的财产状况、风险抵抗能力、还款意愿等相关信息,并作为资格审核与贷款定价的核心依据。〔42〕这些特征已经超出传统歧视特征,选取何种特征及其信贷决策都是基于风险控制的经济理性行为,行为人会承担因决策错误而失去优质客户的风险,为何不能区别对待?“大数据杀熟”也存在类似的论调。“算法不平等”争议的背后反映的是“反歧视原则进入私法领域”,是资源分配过程中公平正义与私法自治的冲突。私法自治与反歧视原则并无位阶上孰高孰低的定论,但算法歧视俨然成为一种广泛的数字风险,对其规制势在必行。为了避免对算法设计者和算法应用者造成过重的负担,需要认真对待规制的时机与程度。

那么,为了平衡私法自治和削弱歧视风险,对“算法不平等”的规制应当选择哪个时机?除了司法救济这一兜底的事后规制之外,事前规制和事中规制成为可选方案。就事前规制而言,“算法

〔39〕 参见许可、朱悦:《算法解释权:科技与法律的双重视角》,载《苏州大学学报(哲学社会科学版)》2020年第2期,第62—63页。

〔40〕 参见陈景辉:《算法的法律性质:言论、商业秘密还是正当程序?》,载《比较法研究》2020年第2期,第126页。

〔41〕 参见李昊:《美国反歧视法治实践中的社会学理论与方法——兼论反歧视诉讼中的统计证据规则》,载刘小楠主编:《反歧视评论》第六辑,法律出版社2019年版,第64页。

〔42〕 参见欧阳日辉、龚伟:《数字信贷、算法歧视与动态竞合政策》,载《南开学报(哲学社会科学版)》2022年第1期,第80页。

偏见”往往涉及民生和公法领域，所以在算法设计阶段或评估阶段等事前阶段就进行强社会法介入。“算法不平等”与之大不相同，面对与日俱增的算法，如果每个算法都需要评估机制和检验机制等事前规制，只会徒增企业成本和公共压力。“算法不平等”的事前规制只适合企业的内部审查，不宜以法律强制。为此，为了尊重市场规律和意思自治，保持科技创新，本文主张事中阶段的规制时机。事中规制是一种过程性规制，其时机为企业将算法应用在用户这一行为过程当中，在行为过程中审视具体的资源分配。本文主张的事中规制方案有二：一是行政监督，二是交易主体间的互动监督。行政监督作为最为常见的事中规制手段，虽孔武有力，但面对不计其数的数字交易行为显得分身乏术。互动监督虽然缺乏强制力，但是将规制成本转移至交易主体之间，同时可以灵活地处理数字技术的不确定性，能够补足行政监督的缺陷。互动监督的核心价值取向在于加强用户方的博弈能力，以对抗涉嫌算法歧视的决策行为。这一点在“算法偏见”中是难以做到的，因为在“算法偏见”中，算法应用者往往是国家机关或能够决定个人重要事项的企业，算法用户只能借助外部力量来对抗。

（二）“算法不平等”的规制维度

对于“算法不平等”，算法歧视现象需要安置在具体应用场景，考察具体的资源分配情况，施以行政监督和互动监督。对于其规制程度的问题，通过下述两个维度来考量：一是通过价值维度来判断算法应用的利益分配和位阶排序，主要施以相应领域法的实质价值约束；二是通过信任维度使算法可信，施以形式平等的规制。

1. 价值维度：通过相应领域法的价值判断施以规制

对于“算法不平等”，首先受到具体应用场景的实质价值约束，然后根据实质价值约束施以相对应的行政监督。相关的实质价值内生于应用场景之中，因为不同的应用场景对应着不同次序、不同权重的价值元规则。〔43〕对于场景内生的实质价值，其内容确认方案有二：方案一为，实质价值来自社会秩序。典型观点为海伦·尼森鲍姆（Helen Nissenbaum）的场景公正理论，其认为社会中存在关于信息及其流动的“活法”，由长期的社会实践和相关情景所形塑，以规范、指导和协调个人信息的流动。〔44〕此理论在算法规制中有不少追随者。方案二为，实质价值来自与场景相对应的领域法。譬如，有学者结合科学路径以如何研究“公平算法”为出发点，认为科学家基于对“公平”（fairness）的不同理解而采取不同的计算公式，背后实则对应着不同的法律概念和法律机制。〔45〕本文青睐于方案二，追求完备的学术理论虽然是法哲学的浪漫与崇高，但同时应当兼顾法律的现实主义需求，方案二相较于方案一更为稳定和具有可操作性。为此，具体“算法不平等”应放置在相应领域法内作价值判断，由领域法提供实质价值约束，行政监督也以此作为介入尺度。

以“大数据杀熟”为例，其实质价值约束主要来自反垄断法，并施以相应的行政监督。“大数据杀熟”作为一级价格歧视，一方面，从市场和社会福利的角度出发，其确保资源分配得到优化，充分践行经济法的有效竞争原则；另一方面，从消费者权益保护法来说，〔46〕其侵害消费者公平交易权。

〔43〕 参见唐林垚：《公共治理视域下自动化应用的法律规制》，载《交大法学》2022年第2期，第24页。

〔44〕 See Helen Nissenbaum, *Privacy in Context: Technology, Police, and the Integrity of Social Life*, Stanford University Press, 2010, p.3.

〔45〕 See Doaa Abu-Elyounes, *Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness*, 2020 University of Illinois Journal of Law, Technology & Policy 1, 54 (2020).

〔46〕 参见胡元聪、冯一帆：《大数据杀熟中消费者公平交易权保护探究》，载《陕西师范大学学报（哲学社会科学版）》2022年第1期；王佳琪：《大数据“杀熟”的法律应对》，载《人民法院报》2019年6月11日，第2版。

仔细思考,“大数据杀熟”自身具有一定的市场调节机制,消费者的地位并非如此低下。由于定价方式以个人需求函数取代全市场需求函数,“大市场”(统一定价的传统市场)被拆分成无数“微市场”(仅有一个卖家和一个买家,卖家需要仔细衡量买家的需求),消费者谈判能力大大增加。虽然单个“微市场”收益微不足道,但杀熟定价模型是统一的,失去一个微市场意味着总体市场失守的可能,因此商家必须放下身段认真研究每一个顾客的诉求,否则就会被消费者“用脚投票”逐出市场。〔47〕另外,消费者也有一定反制能力,诸如通过设置隐私程序迫使厂商吐出更多利润或借助数字比较工具货比三家以避开个性化高价。〔48〕数据表明,大部分算法个性化定价的波动范围为1%～4%,〔49〕属于可接受的一般价格,而非具有危害性的超高价格和超低价格。仅当市场调节机制失衡,“大数据杀熟”造成了经济损害性结果时,法律的介入才是合理的,诸如垄断厂商使用“大数据杀熟”,一级价格歧视的经济优势不复存在,应当适用反垄断法予以规制。〔50〕在这个价值判断下,市场监管部门应采取包容审慎的态度,对差别定价行为给予一定的“观察期”。〔51〕在企业进行差别定价的行为过程中,对于游离在安全线边缘的行为,市场监管部门可以实施适度干预,如对企业进行约谈,要求其对行为进行说明、调整等;对于触碰安全线的,市场监管部门对企业课以相对应的行政处罚。

需要进一步追问的是,对于“算法不平等”问题,在价值判断上是否仅需具体应用场景的内部秩序要求?是否需要额外的实质价值保护?本文的观点是不需要。无论是“算法偏见”还是“算法不平等”,算法歧视的相关争议都来源于算法设计者和算法应用者通过算法权力限制个体的自由意志。根据限制的程度和其所关涉的利益类型,法律的保护力度不同。对于“算法偏见”来说,相关歧视问题造成个体无法融入民生和公法领域。为此,本文对“算法偏见”问题提供“社会融入”的实质价值保护,施以结果义务。相反,对“算法不平等”的法律规制总体持保守的态度。私法领域实际上不存在绝对的平等关系,看似自愿达成的“互利”交易,在其中或多或少存在一方主体基于优势地位对另一方主体“占便宜”,司法对此是克制的,只有到达不公的程度才会干预。〔52〕同时,市场行为往往会针对不同顾客谋划形形色色的差异方案,为此,对算法设计者和算法应用者强加额外的价值要求是强人所难的,并会造成分配正义的泛滥。但这也不是一味遵循具体场景的既有内部秩序即可,否则将会彻底沦为“用旧法律应对新事物”的守旧做法。对于“算法不平等”,现行算法治理对特殊事项仍有少部分额外的价值要求,诸如《个人信息保护法》禁止“在交易条件上实施不合理的差别待遇”,以及对大型平台施以守门人义务。面对“算法不平等”问题,法律虽然不便给予过多的实质价值要求,但可以通过信任维度给予形式平等的保护。

2. 信任维度:算法解释作为互动机制提供形式平等保护

出于算法的黑箱性和复杂性,算法决策的价值维度和信任维度往往是分割的,信任维度不依赖于价值维度。仍以“大数据杀熟”为例,上文虽然已经从价值维度论证了“大数据杀熟”在大多数情况下是合理的,但大数据价格歧视是否为公众接受显然是另一码事。从大众对价格的公

〔47〕 见前注〔2〕,王潺文,第59页。

〔48〕 参见喻玲:《算法消费者价格歧视反垄断法属性的误读及辨明》,载《法学》2020年第9期,第96页。

〔49〕 参见雷希:《论算法个性化定价的解构与规制——祛魅大数据杀熟》,载《财经法学》2022年第2期,第152页。

〔50〕 参见山茂峰、郑翔:《算法价格歧视反垄断规制的逻辑与进路》,载《价格理论与实践》2020年第5期;见前注〔48〕,喻玲文。

〔51〕 参见王文君:《算法个性化定价的反垄断法反思》,载《甘肃政法大学学报》2021年第5期,第154页。

〔52〕 参见许德风:《合同自由与分配正义》,载《中外法学》2020年第4期,第986—987页。

平感知出发,“大数据杀熟”仍然直觉地被视为不合理,立法者也不可能以“专业性”为由,无视大众的朴素认知。鉴于此,不少地方立法已明文禁止“大数据杀熟”,诸如《上海市数据保护条例》《深圳经济特区数据条例》《浙江省电子商务条例》等。所以说,信任维度是规制考量不可或缺的面向。沿着信任维度,法律可以针对“算法不平等”采取多大程度的规制呢?

信任主要通过互动来达成。在算法歧视的讨论当中,歧视与平等问题并非单纯的技术探讨,技术人员与利益相关者的互动能更好地解决问题。〔53〕良性的和有效的互动由确保各方利益的程序来维护,在技术保障不足的当下更是如此。〔54〕算法程序正是以这种方式介入到算法歧视问题当中,但是在不同的算法歧视中扮演不同的角色。程序在“算法不平等”的规制中强调的是互动,以达信任;而在“算法偏见”的规制中强调的是限权。通过互动达成信任这一方案并非要求设计专门的算法程序制度,而是尽可能在规制中体现程序性思维。在“算法不平等”规制的信任维度上,算法解释是关键,并应当附着上程序色彩。由于算法黑箱的存在,算法设计者和算法应用者难以给出确切的解释,算法解释若被理解为“信息控制”的实体性权利将会沦为纸面权利,因此对于算法解释应转向程序性权利的理解,将算法解释视作一种沟通和互动机制。〔55〕作为互动机制的算法解释正是上文提及的主体间互动监督,通过强化用户的博弈能力,提供形式平等的保护,变相对算法设计者和算法应用者施以规制。

算法解释作为互动机制意味着:(1)合作互惠的私法性质使得算法解释应当以追求信任关系为目标。“算法不平等”的主要运用场景以私法为底色,以合作互惠为前提。尤其是许多算法长期与受众打交道,这时候就更有必要建立贝叶斯均衡(互相满意),以达持续性地相互信任,而非单方被赋予过多的解释权能,这样不利于合作的展开。在这个意义上,作为互动机制的算法解释,是一种实用主义解释,关心的是接受者的兴趣以及信任问题。〔56〕(2)算法解释是一种相对权利,算法应用者和接受者结合场景在不同时间段行使和履行不同程度的算法解释权利和义务。一般来说,在算法应用的告知环节,算法应用者(如企业)应当进行概括解释,描述算法的整体运行,让用户感知算法的大致运行状况,帮助个体做出更佳决策,而非披露算法的所有参数和所有细节。但仍有部分受众对算法抱有疑问,受众有权主张个案解释,要求算法应用者尊重个人的知情同意,算法应用者将对算法决策使用了个人的哪些信息并如何做出结论做进一步解释,如果用户仍无法信任可以主张拒绝自动化决策的权利。为此,《个人信息保护法》第四章所规定的信息权利诸如反自动化决策、更正权和删除权等是置于程序性算法解释这一互动机制之下展开的,允许用户在无法被说服的时候因地制宜地行使。对于具有重大影响的算法,算法应用者似乎有必要披露更多的算法信息,要披露多少以及如何披露,以是否被市场驱逐为风险和代价由算法应用者自行判断。

五、结　　论

明确治理对象方能适配相应的治理手段。有学者从技术角度区分不同的算法歧视并提出相

〔53〕参见张贵红、邓克涛:《社会化研究框架下算法公平性的实现策略研究》,载《科学学研究》网络首发2023年2月7日,第9页,https://doi.org/10.16192/j.cnki.1003-20230206.001.

〔54〕See Teresa Scantamburlo, *Non-empirical Problems in Fair Machine Learning*, 23 Ethics and Information Technology 703, 710 (2021).

〔55〕参见丁晓东:《基于信任的自动化决策:算法解释权的原理反思与制度重构》,载《中国法学》2022年第1期。

〔56〕见前注〔39〕,许可、朱悦文,第62—63页。

应的法律规制。〔57〕与之不同的是，本文尝试从制度层面上的性质划分着手，通过区分“算法偏见”和“算法不平等”以明确算法歧视的治理对象，进而提出与之适配的治理手段，以消除治理对象泛化和治理手段碎片化这两个弊端。

这种区分的意义在于：一是从宏观角度来说，其可提供一个融贯的理论解释图景，将多样化的实践问题置于对应的法律框架之中。试举一例：住房福利的算法自动化配分在国外早已有之，〔58〕相关歧视问题固然安置到“算法偏见”来处理。但如果把场景切换成私领域，不排除将来会出现房东和商家通过自动化决策来筛选租客，该场景似乎类似于“大数据杀熟”。但是，住房问题属于民生和个人实现相关的重要领域，理所应当作为“算法偏见”来处理。二是从微观角度来说，其有助于明确具体义务。对于算法设计者和算法应用者来说，他们往往由于被贴上“强者”的标签，而被课以各种宽泛和沉重的义务，这将严重抑制市场活力和主体创新。本文提出的算法歧视的两副面孔明确了算法设计者和算法应用者的具体义务。如果他们的产品涉及“算法偏见”，则主要履行结果义务，根据具体情况适用算法影响评估或“数据控制”，而非算法解释义务；如果他们的产品涉及“算法不平等”，则相关义务主要依赖于算法解释，以变相提高用户的博弈能力。

Abstract There are two flaws in the current legal regulation of algorithmic discrimination, one is that the regulated scope is too board, and the other is that the regulated tools are fragment. The solution comes down to the concept of discrimination in itself. Distinguishing between legal respect and moral discrimination, algorithmic discrimination has two faces. The first face is “algorithmic prejudice” which reflects the discrimination in legal respect and aims to address historical prejudice in algorithmic application, legal regulation of algorithmic prejudice connects with anti-discrimination law, which impose consequentialist liability on the designer and user of algorithm. The second face is “algorithmic inequality” which reflects discrimination in moral respect and aims to address the distribution of resources in the information age. In order to maintain the balance between respecting the autonomy of the will and reducing the discriminatory risk in the issue of “algorithmic inequality”, besides ex-post regulation, the law can enact an interim regulation in value and trust respect to impose appropriate liability on the designer and user of algorithm.

Keywords Algorithmic Discrimination, Algorithmic Prejudice, Historical Prejudice, Algorithmic Inequality, Resources Distribution

（责任编辑：雷槟硕）

〔57〕 参见曹博：《算法歧视的类型界分与规制范式重构》，载《现代法学》2021年第4期。

〔58〕 参见[美]弗吉尼亚·尤班克斯：《自动不平等：高科技如何锁定、管制和惩罚穷人》，李明倩译，商务印书馆2021年版，第71—103页。