算法歧视的成因、认定与治理

许衍会
（南宁威宁投资集团有限责任公司）

摘要：机器学习被广泛应用于自动化决策中，在当前“弱人工智能”时代，用于自动化决策的机器学习模型的性能受制于输入的数据质量及算法设计的优劣，其引发的歧视问题称为“算法歧视”，算法歧视不断暴露，危及公众的基本权利或特定民事权益，亟需规制。算法歧视具有隐蔽性，认定困难，有必要借鉴域外经验推动我国算法歧视认定的标准化；机器学习不同阶段均可能导致最终算法歧视的发生，算法歧视的治理需对从问题定义、训练数据收集到模型训练应用全过程予以规范，对算法歧视的有效治理一方面要求对机器学习及运作过程的深入理解，引入计算机科学中的若干概念、实践规范，对症下药，以技治技、引技入法。另一方面，算法治理需法律规制以保证数据可追溯性与算法可审计性，构建健全完善的公法规范、私法救济体系，以有效治理算法歧视，促进经济社会的公平与正义。

关键词：计算机科学技术；机器学习；算法歧视；法律治理；技术治理；透明原则
中图分类号：TP181; D920.1

Research on Algorithm Discrimination: Causes, Judgment and Governance

XU Yanhui
(Nanning Weining Investment Group Co., Ltd)

Abstract: Machine learning is widely used in automated decision-making. In the current era of "weak artificial intelligence", The performance of machine learning models for automated decision making is limited by the quality of the input data and the design of the algorithm, which gives rise to the problem of discrimination known as "algorithmic discrimination". Algorithmic discrimination is constantly being exposed, endangering the fundamental rights of the public or specific civil rights, and is in urgent need of regulation. Algorithmic discrimination is insidious and difficult to identify, so we can draw on overseas experience to promote the standardisation of algorithmic discrimination in China; different stages of machine learning may lead to eventual algorithmic discrimination, so the governance of algorithmic discrimination needs to regulate the whole process from problem definition, training data collection to model training application. On the one hand, effective governance of algorithmic discrimination requires a deep understanding of machine learning and its operation, and requires the use of a number of concepts and practices in computer science to prescribe the right remedy for the right problem, to treat technology with technology, and to turn technical governance methods into law. On the other hand, algorithmic governance requires legal regulation to ensure data traceability and auditability of algorithms, and to build a sound and complete system of public and private law remedies to effectively govern algorithmic discrimination and promote fairness and justice in the economy and society.

Keywords: Computer Science and Technology; Machine Learning; Algorithm Discrimination; Legal Governance; Technical Governance; The Principle of Transparency

0 引言

随着信息技术的飞速发展，人们迈进数字社会，数据量爆发增长，加之计算机算力的跨越式提升，第三次人工智能浪潮风起云涌，机器学习、深度学习技术取得了一系列重要突破。技术往往是双刃剑，利与弊同时存在，例如基于机器学习算法的自动化决策（下文简称“机

基金项目：无
作者简介：许衍会（1988-），男，公司律师，工学硕士，主要研究方向：数字法学. E-mail: 739083377@qq.com

器决策”）为经营者和消费者带来极大便利的同时，也易引发社会问题[1]。机器决策具有黑箱性，法律滞后、缺乏正当程序约束易导致歧视问题，即学界所谓的“算法歧视”。算法歧视具有很强的隐蔽性，局限于事后救济治标不能治本，近年来，算法歧视规制这一热点问题引起了法学界广泛的讨论，新观点新视角层出不穷，本文尝试更偏重计算机科学的视角去研究算法歧视问题，分析机器学习不同阶段中可能导致算法歧视产生的根源，进而以法治与技治并重的视角针对性地进行治理，以期构建算法歧视多元多维度的综合治理体系。

1 机器学习算法及算法歧视

1.1 机器学习算法及其相关概念

在计算机科学中，算法表示旨在解决某一问题，在有限的步骤内，对输入数据进行变换处理从而得到预期结果的一系列指令的集合。数据是指与拟解决的问题有关联的各类表格数据、图片及语音等信息资料的集合。模型则是在数据的基础上，利用算法在历史数据中搜索得到、利用数据隐藏的模式解决某一问题的特定输出。由此，可以得到一个不太严谨的公示：机器学习模型=数据+算法。目前，主流算法主要指机器学习算法，而机器学习算法又囊括更为复杂的深度学习算法，如 Transformer、VAE 和 GAN 等，深度学习也在不断向前发展，如机器自主学习的元学习（Meta Learning）。现阶段以算法为核心的人工智能仍处于弱 AI 时代，机器学习模型需要通过历史数据进行训练，寻求以拟合特定场景下问题的求解函数，进而在新数据输入时，进行一系列非线性函数变换，输出模型使用者预期的结果，而其中尚无自主的价值衡量。

算法歧视可以认为是机器决策系统受制于训练数据或是算法设计本身的瑕疵，对具有某些特征的特定群体不合理的差别对待，损害公众基本权利或特定权益，违反经济社会公平原则的行为。本文讨论的算法歧视单纯指机器决策过程中产生的错误及其产生的结果实际适用上造成的不合理的差别待遇，无论机器决策结果如何，均会被直接采用，以排除使用者依个人主观意思决定是否采用该结果的情形。

1.2 机器决策的平等盲区

现阶段机器学习模型从训练数据蕴含的客观事实学习特定的模式、规则，进而对新的输入进行判断，无法对相互冲突的价值进行利益衡量，机器决策本质是信息处理活动，通过收集、处理个体基本信息及行为活动信息，将其进行分类或聚类，进而决定后续的处理方式。机器学习算法决策天然地存在平等盲区[2]：首先，输入数据质量决定了算法的可靠性，算法试图基于历史数据给予个体一个客观的量化“数字身份”，必定与现实世界的个体状况存在偏差，主要原因包括：（1）训练数据的完整、真实及时效性方面的局限，或是训练数据样本维度不够，隐含的信息量不足以支撑机器学习，或是本身数据存在歧视，自然导致机器学习失效，或是历史数据对于预测已经失效，后者可以通过迁移学习缓解，但仍无法完全解决；（2）数据对结果有潜在影响，机器学习算法逻辑易忽略因果性并放大相关性；（3）设计者

和使用者不可避免地会嵌入特定的价值目标。其次，机器决策结果具有“马太效应”（Matthew Effect），机器分类决策的内在逻辑是差别对待，其发掘出的模式和相关性用于区分不同的群体。再次，在算法处理个人信息时信息的溢出效应（Spillover Effect）无法避免。算法会基于群体偏好的分析将某个或某些特征作为该群体中成员的公共特征，这实际上是对个体断章取义地描绘或标记，是对信息过度解读的表现。机器决策过程中会赋予特定群体刻板印象，特定群体中成员受到显性或隐性差别对待，导致直接或间接歧视。算法歧视的产生往往源于一些莫名其妙或偶然的关联，给特定群体中的所有个体贴上相同的标签实际上是对个人平等权的贬损[3]。

1.3 算法歧视的种类

算法歧视的具体表现按照损害的利益范畴的不同，可以分为两类，一类算法歧视损害的是公众基本权利，另一类则表现为损害特定个体权益[4]。前者主要指算法损害不特定主体基本权利，权利人不确定，侵权程度未知，例如性别歧视和种族歧视；后者主要是指算法歧视减损了特定群体或个人的特定民事权益[5]。例如数据杀熟，企业尤其是互联网平台掌握着消费者身份、消费经历等海量数据，通过对用户数据的挖掘，获取其喜好及行为模式等特征，赋予其特定标签，进而制定差异化的营销策略，即包括个性化营销定价，改变了“同物同价”的传统定价模式，这即是所谓的算法杀熟。后者与前者相比具有自身的特殊性。前者损害的利益主体不特定，损害的利益具有普遍性，而后者则是对特定群体的特定利益的损害，主体明确，利益也具有特定性；此外，前者的规制法理基础一般为公平原则，而后者的规制则主要通过具体的部门法来规范，例如《消费者权益保护法》、《个人信息保护法》等。

2 机器学习过程中歧视产生考

2.1 机器学习及训练过程

2.1.1 定义问题与算法设计

利用机器学习解决现实中的问题首先需要将拟解决的现实问题转化为计算机问题，确定是分类、聚类问题？或是离群值检测、生成等问题？只有定义了问题才能进行下一步的算法设计。算法设计是一个非常复杂的问题，当前不同领域采用的机器学习方法各异，例如在表格数据处理方面，往往采用 XGBoost、AdaBoost 等集成算法，而在图像语音领域则采用的多是 GAN、transformer 等深度学习算法，现实中的问题的解决与机器学习算法并非一一对应，例如在信贷风险识别问题上，即可以采用分类学习算法，亦可以采用离群值检测算法。算法的设计是具有高度专业性的工作，算法的确定依赖于设计者对待解决问题的理解、对机器学习专业知识的储备、技术水平能力以及拟应用领域的行业知识等。

2.1.2 数据收集

数据决定了模型的上限，这是计算机科学界对数据之重要性的共同认知。训练数据的来

源包括内部收集和外部获取，以互联网平台为例，消费者网络上的注册、浏览、消费等行为会为互联网公司提供大量的内部数据，此外，互联网公司为了更加全面的去分析客户，可能还会通过爬虫、有偿购买等方式获取更多客户的数据用于其决策。获取的数据依据其是否经过技术处理，可以分为原始数据、脱敏数据和加密数据，原始数据是模型训练最理想的数据，是用户行为的直接体现，而脱敏数据是对原始数据的匿名化处理，往往也可以通过算法学习到数据中隐含的模式，与数据脱敏有本质差异的一种数据处理方式是数据加密，加密数据通常不可用。从计算机科学的角度，数据收集需要考虑多方面的因素，一般样本量越大、维度越高、准确性越高且时效性越强的数据更有利于模型训练，训练出来的模型性能更优。《个人信息保护法》从信息保护角度，规定了“限于实现处理目的的最小范围原则”，虽我们可以理解样本重要特征对结果的影响程度满足正态分布，但是对于此原则下数据收集范围的界定的确存在诸多难点，而且与机器学习中期望更大量更多维训练数据的需求相左。

2.1.3 数据预处理、特征工程与数据分割

机器学习过程中，在数据输入算法之前需要对数据进行预处理，包括数据清洗、数据标准化、数据分割等。数据清洗主要对缺失值和异常值的处理，数据标准化主要是对不同数量级的特征进行归一化处理，以上数据预处理都旨在避免异常样本或某些特征对训练的负面影响。特征工程是机器学习重要的环节，其困难之处不在于技术，而是跨专业的经验，特征工程主要包括特征构建和特征选择，前者的主要做法有特征组合、特征拆分及外部关联特征，其中特征组合通过将现有特征进行组合的方式得到新的、更具价值的特征，特征拆分将现有某个特征拆分成多个特征，一方面可以新增信息表达的维度，另一方面拆分出来的特征可以再与其他特征进行组合，外部关联特征是拓展特征维度的一种做法，旨在引入更多表达信息的维度，例如，对于日期特征可以关联到温度、湿度等外部特征。后者可以通过特征的取值分布、特征与目标的相关性等方法进行特征选择，判断特征与目标的相关性即判断特征间的变化是正相关或负相关，常用信息熵增益、皮尔逊系数等方法，其也是特征选择是最为常用的做法。数据分割旨在从数据集中划分出一部分数据用于模型的验证，以检验经训练集训练而达到收敛的模型在未知数据上的表现。

2.1.4 模型训练与测试

模型训练前需要进行参数初始化，在训练过程中进行算法优化，模型训练一般以算法收敛或者指定训练轮数作为终止条件。训练完成后，使用验证集进行模型验证，确定模型是否符合预期，经训练的机器学习模型可能出现欠拟合或过拟合问题，需要根据具体情况进行模型的重新训练，欠拟合往往需要补充训练样本的数据量或数据维度，过拟合则往往更加复杂，需要从算法和数据两方面进行思考，或是算法复杂度不够，无法有效拟合数据，则需要提升算法的复杂度，例如增加神经元个数，网络层数等。或是数据训练次数过多，非关键特征主导了模型输出，因而需要减少训练的次数。

2.2 机器学习不同阶段导致歧视产生的根源

机器决策过程的不同阶段涉及不同类型的歧视，导致机器决策产生歧视的原因或是由于决策系统本身的错误，例如通过有偏见的数据库训练的决策系统，或是在决策系统的训练期间没有产生错误，然而系统产生的结果对具有受保护特征的群体产生明显不合理的差别待遇[6]。根据上述机器学习过程，机器决策过程分为三个阶段：事前阶段，即数据收集、处理和输入机器决策系统之前的阶段；事中阶段，其机器决策运行阶段，此外，将算法设计也纳入此阶段进行讨论，则事中阶段聚焦算法的设计、运行；事后阶段则是结果输出阶段，即可能产生算法歧视的实害阶段，以上不同阶段，每个阶段可能产生的歧视类型或方式不同。

2.2.1 事前阶段

这一阶段的歧视往往源于训练数据潜在的偏见。训练数据的质量对最终模型的性能起着重要作用。数据承载着社会的价值观念，或先进良好或腐朽落后。模型无法绕过数据原教旨主义，至多削弱数据中的偏见，但无法消除数据中既有的歧视。此外，算法歧视也可能由数据本身的缺陷引起，例如数据不准确和资料不适当。数据不准确是指输入的数据并非训练模型所需要的数据，产生的结果必然也无法达到使用者的期望，同时，由于数据本身的偏差将进一步地导致产生歧视性的结果。数据不适当是指特定群体的代表性不足或过多、非法数据采集或不当的数据关联，则可能产生歧视的形态包括：数据类别不平衡导致对特定群体的过度监管、使用人体特征的不当分析、使用非法收集的数据导致对特定群体的偏见。此外，数据收集和分类方式的差异也可能导致算法歧视。在数据进行标签分类时，需依靠使用者将其进行分类，当使用者在先前决策已受到偏见的影响，而又以该决策结果作为模板时，后续的决策系统即可能推论出具有相同偏见的规则。在数据收集中，也可能受其他因素的影响导致收集的数据过于片面或不准确，更有甚者，使用者可能有目的性地收集对特定群体不利的数据。

2.2.2 事中阶段

事中阶段涉及算法设计和机器决策运行阶段。在算法设计中，设计者和使用者的偏见是导致歧视产生的根源，其将客观事实量化为不同的数值，个人主观意识隐含的偏见很可能移植至机器学习模型。机器决策运行中歧视的产生主要来自两个方面：一方面，特征变量的选择。决策系统对特征变量的采用标准过于粗糙，决策系统采用的特征变量即分类标准与该决策系统的使用目之间关联性低，因此难免有可能造成统计歧视，使得错误的决策结果在不同群体间产生不成比例的情况，要解决这一问题通常涉及决策成本的增加，而使用者大多不愿意花费额外的大量成本以获得更完善的数据，以建立更为准确的标准。另一个方面是代理变量和受保护特征之间的相关性。若代理变量与受保护的特征具有高度的相关性，决策系统以代理特征作为决策的依据时，可以认为决策系统实际上也在使用受保护的特征作为分类依据，极有可能对群体中的个体造成不合理的差别待遇；此外，也有可能发生使用者通过寻找关联性更弱的代理变量来掩盖该决策系统具有歧视性的事实。在这一阶段产生的歧视还包括

后天学习引入的偏见，模型在利用新数据不断训练迭代中，新数据存在的偏见会让原本良好的模型产生偏差，在持续的决策结果中产生歧视，进而决策系统的实际应用会导致不合理的差别待遇。

2.2.3 事后阶段

事后阶段是指在结果输出阶段可能产生的歧视，即是针对决策系统产生的结果进行法律上评价，以确定其是否导致受保护群体中成员受到不合理的差别待遇。在法律评价中可能遇到的最大问题是，当该结果只是统计上必然的结果时，是否应被视为构成歧视。有必要基于实质平等观的立场，从整体社会背景的角度，分析该差别待遇是否牵涉到社会对特定群体长期存在的歧视性对待。

3 算法歧视违法性判断

3.1 不同平等观下的歧视

不同平等观点下的歧视可分为形式平等下的歧视和实质平等下的区别，有必要区分形式平等和实质平等[7]，具体如下：（1）形式平等追求所有人之间的无差别待遇，其核心观点强调人人平等。除了承认平等权是一项个人权利外，也注重法律上的平等，即司法平等，即要求首先根据现有法律规定判断是否存在禁止差别待遇的情况，如果答案是肯定的，则进一步审查差别待遇的手段是否具有合法目的且不具可责性。如果不存在“法律上”的差别待遇，而只涉及“事实上的地位”不同时，形式平等的保障就失效了。形式平等忽略外在社会、文化以及经济环境对于弱势群体造成普遍性和结构性的不合理的差别待遇。（2） 实质平等强调，除了不同群体在不同权力关系中的对等地位外，更要求将平等权利扩大到“事实上”或“社会关系”，而不仅仅是法律地位上的平等。实质性平等要求国家积极消除对受到歧视的弱势群体的不合理的差别待遇，使他们获得与其他群体实质上相同的地位，从而允许国家利用不对称的差别待遇来实现这些目标。法律上的平等或事实上的平等是从不同角度进行的具体评价，前者是一种法律评价，因为评价标准是法律，而后者是从社会权力不平等关系出发评价社会上的事实。形式平等观将平等权定义为个人的权利，并将差别待遇的概念限制在法律背景下，忽视了外部结构性因素。而实质平等观除了以群体的保障为基础之外，其差别待遇的认定标准主要基于社会事实上的平等与否。在 2021 年出台的《个人信息保护法》中，立法者试图通过若干权利和义务分配规则来纠正主体之间资源和能力不平等的问题，以达到实质化平等的目的，避免个人信息的过度收集和滥用、算法黑箱等问题，实现对个人信息权益的法律保护。然而，《个人信息保护法》实际上只提供了一个框架条款，有待进一步细化。

3.2 直接歧视与间接歧视

3.2.1 直接歧视的内涵

直接歧视指故意对具有某一特征的个人或者群体相较于其他具有类似地位的群体予以

不合理的差别对待的行为，且差别待遇的原因是由于受歧视的个体或群体的种族、肤色、宗教、性别或国籍。综合域内外法律，直接歧视的要件大致上包含受法律保护的特征、被不合理差别对待的事实、存在主观故意及前两项之间存在因果关系。对是否构成直接歧视的判断是相对直观的，决策行为是基于使用受法律保护的特征，进而对特定群体实施了不合理的差别待遇，导致其权益受损，且在这个过程中存在算法使用者的主观故意。（1）受法律保护的特征。该要件要求主张受侵害的人是具有受法律保护特征身份的人，以我国《宪法》为例，其中明确受保护的特征包括民族、宗教、性别等基础特征，《宪法》第四条中规定：“中华人民共和国各民族一律平等”、“禁止对任何民族的歧视和压迫”。第三十六条中规定：“不得歧视信仰宗教的公民和不信仰宗教的公民。”第四十八条中规定：“中华人民共和国妇女在政治的、经济的、文化的、社会的和家庭的生活等各方面享有同男子平等的权利。”上述并未涵盖所有受保护的特征，其他部门法中仍有明示的受保护的特征。此外，法院也需要通过对法律条文的解释将受保护的特征范畴扩张解释到其他未明示的特征，如是否已婚、工作单位性质等特征。（2）被不合理差别对待的事实。该要件是指原告如果要主张受到直接歧视的侵害，首先需要对其受到不合理的待遇对其实体上权利的影响或受到的损害进行举证，理论上来说构成不合理的差别待遇的规则、政策、法规能够从相关文件中确立，对于是否构成不合理的差别待遇，难点在于“和谁比较”作为确认原告受到不合理的待遇的依据，学界也提出在难以找到比较基础时采用假想的比较对象。（3）存在主观故意。该要件根据算法使用者主观意图进行判断。有学者提出以主观上有无偏见和客观上有无歧视进行区分，当存在主观偏见且歧视客观发生时，构成直接歧视；如果没有主观偏见，但歧视客观存在，则构成间接歧视；无论是否存在主观偏见，只要客观上不存在歧视，就不会构成直接或间接歧视。因此，在确定算法使用者是否存在主观意图时，有必要确定歧视是否客观发生，而主观偏见的存在与否只影响歧视的类型。（4）因果关系。该要件要求前两项（1）（2）之间存在因果关系，学者提出多种判断标准，有条件说、原因说、法规目的说、盖然因果关系说及相当因果关系说。相当因果关系说不要求行为与损害结果之间具有直接因果关系，以行为所造成的客观事实作为基础，在确定行为与结果之间有无因果关系时，依行为发生时的一般社会经验和知识水平作为判断标准，判断该行为是否可能引起该损害结果的发生，实际上该行为又确实引起了该损害结果，则认定该行为与结果之间有因关系。冯·克里斯认为，主张行为与损害之间具有相当因果关系，必须符合二项要件：该行为必须是损害发生的必要条件；该行为必须一定程度增加了损害发生的客观可能性。相当因果关系说扩大了因果关系的范围，使算法使用者不会被轻易免除责任，有利于受害人利益的保护。作为大陆法系的通说，相当因果关系说具有两点明显的优势，一是有机结合了“条件判断”和“相当性判断”。二是将“可能性”纳入因果关系判断中，受害人的利益可以得到更好的保护。相当因果关系说在判断行为是否导致损害后果时，只需证明该行为极大地增加了损害的可能性，而并不要求对因果关系的证明达到精准和绝对的地步。因此，相当因果关系说减轻了受害人的举证责任。因此，对于直接歧视，采用相当因果关系说来确定因果关系更为合理。

3.2.2 间接歧视的内涵

美国法律将间接歧视定义为表面上中立且无歧视故意的规则或行为，但在实际应用中会对群体中成员产生不成比例且不合理的影响。间接歧视是否要求算法使用者具有歧视性目的或意图，这在美国法律中存在分歧。《宪法》要求原告证明算法使用者具有歧视的目的或意图，作为确立间接歧视的必要条件；但美国民法则并不要求证明主观目的或意图的存在，即表面上中立的规则或规定导致对具有受保护特征的群体成员的明显不合理的差别对待时，即推定存在间接歧视，此时举证责任倒置，由决策方证明其行为或规定未以受保护的特征作为分类标准，而产生差别待遇的结果。欧盟法律和《欧洲人权公约》将间接歧视定义为算法使用者表面上采用中立的区分标准或普遍适用的规则，然而其结果是，具有受保护特征的群体成员相对于其他具有类似地位的其他个体而言，将受到“明显不合理”的影响，且上述事实上的不合理结果产生的间接歧视无正当的理由予以排除其违法性。此外，欧洲人权法院和欧洲法院都认为主观上的意图并非构成间接歧视的要件。

间接歧视的判断较为复杂，其不要求算法使用者有主观上的目的或意图，且算法使用者利用的规则表面上是中性的标准，因此需要通过结果判准认定算法使用者的规则是否构成间接歧视。构成间接歧视的要件包含：表面中立的规则、因果关系、用作比较的对象范围以及判断该差别待遇是否明显不合理，以下将针对每个要件分别进行讨论：（1）表面中立的规则。表面上中立是指相关规则并非未受到法律明文限制，且对所有具有受保护特征的群体均普遍适用。规则是否属于表面上中立的判断标准是通过算法使用者的主客观以及其对于具体规则有无控制权进行判断。（2）因果关系。与直接歧视的因果关系要件类似，不再赘述。（3）用作比较的对象范围。即法院判断原告是否被歧视中用作对比的观测对象。该要件认为法院观测的对象不仅以实际符合资格的人受到的影响为基础，而且也应从相关规则对于整体数据库的影响进行分析。当原告声称存在差别待遇时，一般以实际申请人或其他符合标准的人受到的影响作为证据，此外，也有观点认为应通过与个人相关的数据库受到的影响来证明算法使用者的规则存在差别对待的情况，前者的优点在于样本母数少，法院更容易审查，并且不容易出错或存在操作不清楚的情况。缺点是忽略了样本在当地人口中的估计比例，如果比例太低，就有可能代表性不足。此外，将法院的观测对象局限于这类样本可能会忽视少数群体长期存在的结构性不合理差别待遇地位。后者则是以区域内人口中具有该受保护特征者的比例作为观测的对象，并根据该少数群体长期以来受到的结构性不合理的差别待遇，评价算法使用者的规则是否构成差别对待。采用这种样本类型更符合上述实质平等的内涵。法院应允许原告通过社会层面的证据证明差别对待的存在，而无须判断因算法使用者采用的规则受歧视的人是否有足够的数量，此外，间接歧视本质上是关注具有受保护特征的群体受到事实上不平等待遇的情况，且间接歧视规制的目的是为对抗社会整体结构上产生的结构性歧视。因此，法院在判断是否构成间接歧视时应将整个群体作为判断的范围，以达到间接歧视规制的目标。（4）构成明显不合理差别待遇的判断方法。美国和欧洲法院对间接歧视的认定分别采用“不成比例”的认定方法和“明显不合理”的认定方法。法院判断算法使用者的

规则造成的差别待遇是否构成明显影响时，美国法院除适用美国公平就业机会委员会（EEOC）对于统计数据提出的标准外，也有部分判决采用其他针对统计数据的标准；而欧洲法院除了使用数据统计标准进行判断外，提出数个“质”层面的判断方法。

当法院认定算法使用者采取的表面中立规则对具有受保护特征的群体造成差别影响时，并不会直接认定该规则侵害特定群体或个人的平等权。算法使用者可以通过主张其采取该规则具有正当化的事由，以避免违反禁止歧视的规范。有差别影响事实的存在只能推定相关规则有构成间接歧视的可能，在其具有正当化事由的情况下，即使规则造成一定程度的差别影响，但该差别对待失去可责性。在美国法体系下，法院是针对规则是否具有必要性进行审查，若具有必要性，算法使用者可据此主张其具有阻却违法事由；在欧洲法院的体系下，法院则是通过规则是否具有正当的目的，以及其手段和达成目之间是否合乎比例原则作为审查的标准。

3.3 算法歧视的违法性判断方法

3.3.1 直接歧视的判断方法

直接歧视以算法使用者有歧视的故意为前提，算法设计者主观上以带有歧视的方式构建自动决策系统，算法设计者的歧视意图被移植到机器决策中，进而“间接影响”决策结果。以下三种情况可认定具有歧视性的意图：运用受保护的特征对群体进行区分时未审慎考察、使用有问题或受到污染的数据、完全依赖模仿人类先前歧视性行为或造成偏见的机器决策。并非所有的歧视性意图皆须禁止，而是应判断该歧视性意图导致的决策是否达到“贬低”受保护群体的程度，造成受保护群体的成员丧失个人尊严或无法获得他人的尊重。“贬低”这一概念的理解从客观和主观两个角度进行讨论。从主观上，决策系统基于算法作出决策，核心问题在受不合理差别待遇的受保护群体成员得否认识到，机器决策对其做出的决策存在对其人格的贬损及不尊重。然而，公众普遍难以感知；客观上，算法使用者的行为是否达到“贬低”的程度不应仅从个人主观的认知进行判断，而应同时评价算法使用者实际上的作为或不作为义务，仅当算法使用者能够采取适当的措施限制歧视性结果发生而未采取，且发生歧视性结果是可预测时，才能认定算法使用者的行为主观上具有“贬低”的涵义，而有认定该决策系统造成歧视的可能。直接歧视的构成不要求算法使用者有歧视的主观意图，因此有必要讨论算法使用者在没有歧视意图的情况下使用机器决策产生对受保护群体歧视行为的情况。学界认为，即使客观上受影响的人可能不知道他们正在受到决策系统的歧视，只要对决策系统使用者的作为或不作为进行审查，如果其能够采取应对的措施却未为之，进而造成歧视性的结果时，即应认定其具有歧视性意图，有构成直接歧视的可能。

3.3.2 间接歧视的判断方法

间接歧视的前提条件是“表面中立”行为，直观上很难认定这种表面中立行为是否会对受保护群体造成“法律上重大”或“实质性”的不适当差别待遇，有必要使用“间接歧视”的审查标准作为法院裁定的依据。在实践中，欧洲法院和美国最高法院提出了不同的方法来

确定表面中立的行为是否构成间接歧视。在介绍相关方法之前，需要界定两个概念：有利群体和不利群体，有利群体可以狭义定义为在特定场景中，在某一决策行为作用下，拥有同样某种受保护特征的参与群体中个人基本权利得到保护且个人权益得到增加的群体。不利群体则是上述特定群体中剔除有利群体成员以外的人员构成的群体。欧洲法院提出了“人口统计（差异）平等（Demographic (Dis)parity）”和“显性负面作用（negative dominance）”的方法，这两种方法都是以具有受保护特征的群体在有利群体 P_1 及不利群体 P_2 中估算的比例进行比较。前者以全体人员的比例进行计算，计算结果可能受到其他外部因素的干扰，但相对而言是一种较为合理的判断方法；要构成后者设定的标准，除了 P_2 须大于 P_1 外，P_2 本身亦须大于 50%，这一额外的条件，导致反歧视法对基本权利的保护取决于具有受保护特征的群体的规模。

此外，美国联邦最高法院则提出了“4/5 规则”、“2 或 3 个标准偏差分析”以及“多元回归分析”等判断方法[8]，“4/5 规则”是指使用者做出的决策结果中，任一群体的通过率，若低于通过率最高群体的 80%，即构成明显的差别待遇，该规则的缺点在于其仅能适用于与选任有关的案件，且判断方式过于僵化。“2 或 3 个标准偏差分析”则是指当决策结果未落在正态分布的区域内时，可认定该结果不是随机产生的，应被视为构成明显差别待遇的情况。而“多元回归分析”是对自变量和因变量之间是否存在相关性的研究，法院应以各种不同的自变量进行测试，并使用测试结果来确定自变量与因变量之间是否存在统计上显著的关系，以及它们是正相关还是负相关。此方法并不能确保与因变量具有统计上显著关系的自变量是为受保护的特征，并且测试成本极高。

有学者提出的以“附条件的人口统计差异”（Conditional Demographic Disparity）作为算法决策结果是否构成明显不合理差别待遇的依据。设受保护特征 X，则具有 X 特征的有利群体（不利群体）中具有受保护特征的人数比例 P_{X1}（P_{X2}）即为具有 X 特征的有利群体（不利群体）中具有受保护特征的人数与具有 X 特征的有利群体（不利群体）人数之比。当决策结果不存在间接歧视情形下，P_{X1} 与 P_{X2} 相等；当决策结果存明显不合理的差别待遇时，具有 X 特征的群体中，P_{X2} 大于 P_{X1}。而此种判断方式与前述“人口统计（差异）平等”判断方法的差异在于其添加了额外的条件，将影响法院判断的要素限缩在具有相同条件的人数上，进而可排除统计人数过于广泛而产生不精确结果的情形。附条件的人口统计差异判断方法下，法院通过查看每个案件中涉及的政治和社会因素，确定受不合理对待的群体及与其进行比较的群体中成员组成是否为具有受保护特征的群体。当法院梳理出案件中哪个群体为该案件中具有受保护特征的群体后，就可以通过上述“人口统计（差异）平等”的判断方法，评价该案件中是否存在对具有受保护特征的群体成员造成明显不合理差别待遇的现象。可见附条件的人口统计差异可以处理算法决策下公平性的两个关键问题，一是法院可以确定使用哪些受保护的特征作为比较的对象，以确定决策系统中的潜在歧视。二是当认定作为比较基础的群体后，应如何对这些具有受保护特征的群体的系统输出结果进行比较。然而，附条件的人口统计差异的有效性不应被夸大，即使能通过该判断方法得到在符合特定条件下，具有

受保护特征的群体成员各自占有利群体及不利群体中的比例，单就其判断结果，无法确定是否存在不合理差别待遇以及该差别待遇是否合理。采用附条件的人口统计差异旨在为协助法院、算法使用者等在判断上述问题时，能够有一致的标准作为证据，有助于加害人或被害人减轻其举证歧视存在或不存在的责任。

本文认为应采取学说上提出的附条件的人口统计差异判断方法，即在“人口统计（差异）平等”判断方法的基础上，增设一特定条件，并以符合该条件下有利群体及不利群体中具有受保护特征成员的比例进行比较，此方法不会因具有受保护特征的群体规模过小而无法保障其权益。在进行算法歧视违法性认定时，除了以上内容，仍需考虑算法使用者形式上存在直接或间接歧视，但属于商业习惯和行业惯例的行为，允许算法使用者即使实施了“算法歧视”的行为，也应允许算法使用者进行“正当理由”抗辩。例如，基于平台公平和非歧视规则的随机交易、新用户的促销等行为。

4 算法歧视的多元治理思路

算法歧视是数字社会下歧视一种新的表现形态，首先需要明确的是算法歧视的消除，最关键的角色是算法设计与使用者，其作为直接掌握算法的主体，有义务保障用户的正当权益，秉持良好价值观和商业道德的算法设计与使用者应主动拒绝算法滥用，做好算法的具体内容审核，对应用场景的功能做价值导向判断，对算法的原理做合理性分析，对算法可能存在的自身缺陷做风险评估防护，把风险拦截在萌芽状态，建立算法设计与训练的安全审查，算法测试与上线的稳健性监控，以及日常运行的风险感知及记录存档，积极消除其算法可能对特定群体的侵害，例如，对于推荐、检索类算法，需要加入内容打散等干预功能，并优化规则的透明度和可解释性。

4.1 算法歧视治理的指导原则：透明原则

4.1.1 算法透明的概念

《个人信息保护法》规定了自动化决策的透明度要求以及处理结果公平合理，可见透明原则是法律回应的各种路径当中最直接、有效和恰当的方式。《个人信息保护法》第二十四条第一款提出“决策透明”，虽然透明原则一直是用于规制公权力的一项原则性要求。而算法本身并不是由公权力机关独享，但算法可以自我训练、自我学习，具备动态的规划能力，不再需要人为的干预，部分权力由人转移向算法，形成了学术界所谓的“算法权力”。机器学习算法具有专业性，算法工作原理超出一般公众的理解范围，因此被称之为“算法黑箱”。透明原则最终的落脚点，是对于机器决策的规制。理论上，若算法透明则整个决策和执行过程可知，算法便更容易被监督，也可以消解公众对机器决策提出公平性和合理性的质疑。

学者们对算法决策中透明原则的认识存在差别，普遍认同的最大公约数：“算法的设计方或者使用方对包括源代码、输入数据、输出结果在内的算法要素的披露。”此外，还有学者认为应在此基础之上补充对机器决策是如何做出的解释。以上对算法透明原则的界定，暂

且称之为狭义算法透明原则。狭义算法透明原则存在诸多弊端，例如，有损商业秘密制度的效能、为不法分子创造规避算法的机会、导致因数据公开引发的个人隐私信息泄漏、危及个人信息安全等诸多问题。此外狭义算法透明效果有待商榷[9]，算法代码多为句法、表征、符号等技术语言，公开算法代码不意味着完全理解算法的工作原理和决策过程。从公开的算法代码和数据等信息中理解或核验算法过于理想化。机器决策运行是一个动态的过程，数据不断投喂、算法不断迭代，诸多问题及漏洞需要在算法执行过程中，才能被发掘并加以解决。此外，仍然无法保证排除第三方干预，从而影响最终结果。

4.1.2 算法透明原则重构：从“算法透明”到“决策透明”

2017 年，美国计算机协会公布了算法治理七项原则知情原则、访问和救济原则、可问责原则、解释原则、数据来源处理原则、可审计原则、检验和测试原则。从中可以得到关于透明原则的教益：计算机工程师对直接公开源代码持保守态度，认为公开源代码不等于知情，更应关注“算法设计、执行、使用过程中可能存在的偏见和可能造成的潜在危害”。在算法透明的功用方面，最终运算结果无法通过公开和披露算法准确得到，只能力图接近，但很难确切把握。算法披露本身也受到隐私保护、商业秘密等其他因素的制约，其与算法透明的可行性有着持久的张力[10]。尽管限制披露对象可以缓和这种张力，但这也无法根本解决所有冲突。

对于透明原则的路径重构，思路是从“算法透明”到“决策透明”。机器决策中算法透明的实现需要跨学科、跨领域的专业群体通过对算法的技术原理和社会效果进行匹配对齐，解构机器学习算法决策系统，达到算法透明之目的，我们应该关注的是算法歧视形成的机制，进行体系化的规制。算法透明原则的重构是基于算法问责机制，问责机制的前提是对机器学习决策过程中不同阶段予以法律规范，构建法律规制的界限和依据，以对算法的设计、执行和使用各个环节都具有法律规制力和可责性[11]。2021 年 3 月 1 日，我国“互联网信息服务算法备案系统”正式上线运行，算法备案作为多维一体算法监管体系的关键环节，将与算法评估、算法检查一道成为我国算法监管的“三驾马车”，初步构建起我国的算法监管制度架构，为算法问责机制的实现奠定了基础。决策透明的实现即使在机器学习算法决策过程的不同阶段需要进行有针对性的制度约束，满足合法性要求，需要分别面向政府、公众、商户和用户去进行说明和告知，并把他们的反馈意见引入到算法的更新和迭代之中，来保障各个主体的权利。向不同的主体进行披露以建立这些主体对于算法的信任和理解；在算法的更新和迭代的过程之中，接受外部的反馈、社会的评价以及消费者的建议，让这个算法更好地服务于每一类主体，满足不同主体对于算法的期待，产生更好的社会效应。

4.2 算法歧视治理方法及路径

从治理的方法维度可以将算法歧视的治理方案分为法律路径和技术路径[12]。法律路径指通过法律规范对算法歧视产生过程及结果中存在违法的行为进行规制及救济。例如，《个人信息保护法》对个人信息及敏感数据的收集使用的限制、对造成算法歧视违法行为的算法

设计者和使用者的追责制度、对算法的使用者附加提供相应的算法解释的义务等等。算法审计的法律机制是限制算法歧视的最有效的方式，借鉴安全工程的理论，可以通过预处理方法、内处理方法和后处理方法来进行算法审计。算法审计在我国算法治理中意义重大，可视为算法透明的本土化方案，算法审计作为一种事前的治理方式，也为事后规制提供了条件。而算法备案制度是算法透明的本土化治理方案。目前全球主要经济体采取事前备案方式的只有欧盟和我国。欧盟《人工智能法案》中对备案的内容、程序等均缺少具体规定，而我国在积极推进细化算法备案的具体规则、法律后果，构建算法监管之“地基”，现阶段我国算法备案制度仍在不断发展与完善，例如对于算法备案的义务主体，未来应当进一步扩大。就算法备案的范围而言，应当采取适应算法技术发展的分类分级方式，向“工具类”和“智能类”的分类和更着重强调风险的分级方式转变。立足于算法监管的全流程，应当细化相关规则，促进企业履行备案义务。在算法监管的全流程中，应当对算法初始备案和动态备案规则进行细化，从而平衡算法安全和算法发展。技术路径强调算法歧视是“传统歧视”经过“算法”复制或放大后形成的产物。消除算法歧视的唯一的解决方案就是消除传统歧视本身，通过对数据进行预处理，消除其中隐藏的偏见；或删除敏感属性及其统计相关的代理属性；或在算法中添加随机性以避免过拟合及隐藏算法偏差；或对最终的判别模式进行审核，来排除可能带有歧视的模型；或利用添加“正则化项”约束算法修改目标函数，寻找可以消除歧视的模型。

治理的时间维度方面的研究以古德曼等人提出三阶段划分为代表，其以一种更加技术化的方式对三阶段的算法治理进行了描述与解读，分别为以数据集处理消除歧视的前处理阶段、以修改模型形式或结构治理歧视的内处理阶段和以改进模型的输出结果消除歧视的后处理阶段。前处理的方案属于事前治理，尝试消除作为输入的数据中潜在的歧视，包括为改变历史数据中记录的最终决策对分类超平面边缘的标签进行重标记、为更改歧视性数据在生成模型中所起作用的重要性对数据的权重进行修改、为尽量避免使用可能含有歧视的数据而选择性地使用数据对数据进行重采样等方法。内处理阶段和后处理阶段则将关注点聚焦在了算法上，均属于事中治理，二者的不同主要在于内处理应用在训练阶段，后处理对应的是测试阶段。对于内处理阶段，古德曼提出在算法训练中设计一个“公平正则化项”来约束模型训练，通过优化算法的不断迭代来予以修正，阻止算法歧视的产生。对于后处理，主要内容是模型训练完成后到部署上线前，利用测试集数据，对模型的输出结果进行改进，防止实害的产生。

本文在古德曼等人研究的基础上，结合本土治理的经验和做法，将算法治理的时间维度划分重构为以数据治理为中心的事前阶段，以算法治理为中心的事中阶段和以救济保护为中心的事后阶段，事前阶段指模型训练之前，收集数据、处理数据，完成特征工程的阶段，事中阶段指算法设计、接收数据后利用数据训练算法并生成模型的阶段；事后阶段指模型部署到程序之后对算法应用造成损害后的事后救济，是算法歧视的实害阶段。

4.2.1 事前治理：以数据治理为中心

机器学习模型的训练需要大量的数据，在数据层面加强保护是规制算法的重要一环[13]。

事前治理主要包括：对数据的获取加以控制、对数据进行预处理、建立退出权、保证数据的可追溯性。（1）数据收集的约束。即限制收集或使用数据的类别和数量。例如，《个人信息保护法》要求在收集个人的数据时征得个人的同意，或者增加其他的限制约束数据的收集。《民法典》第 1034 条规定了自然人的个人信息受法律保护的权利，随后《个人信息保护法》在对个人信息收集、处理和相关责任的问题上做出详细规定，其中的合目的性、最小化、透明性和安全性原则使其成为数字时代个人信息保护的“基本法”。欧盟《通用数据保护条例》对机器学习中数据的收集应满足“合法性、合理性、最小化、准确性、透明性和有限性”。（2）数据预处理。对数据的预处理是对原始数据的二次加工，是对算法歧视在数据环节的针对性的治理，旨在消除可能导致算法歧视的数据的影响。具体可分为以下三点：一是删减数据维度。直接删除收集的数据中的敏感属性和代理属性下的所有维度，排除具有明显歧视特征的数据维度，例如种族、肤色等特征；二是删减部分数据。对可能涉及歧视的“异常”数据进行删减，关键在于历史数据中带有歧视性的样本的识别确定，避免对模型训练的干扰；三是更改抽样的方式。通过改变采样的方式来降低歧视发生的概率。（3）建立“退出权”。维克托·迈尔－舍恩伯格在其《删除》一书中提出，人类应该具有删除其在互联网上留存的数据，进而被互联网遗忘的权利。数字时代的人不能够将自己排除在数据海洋之外，至少应该有拒绝成为算法客体的权利。一方面赋予公众匿名化的权利。匿名化并非指真实姓名的匿名，而是作为一种数字人格的匿名。匿名化处理要求算法使用者对用户去标签化，不对其进行画像建模和特征分析，匿名化是保持数字时代个人安宁权的有效方式；另一方面从法律上保障“退出权”的建立。欧盟《通用数据保护条例》中率先提出了用户有免受完全自动化决策对待的权利，我国《个人信息保护法》也规定，个人有权拒绝处理者仅凭自动化决策作出对个人具有重大影响的决定，这为退出权的建立提供了依据。退出权的内涵是用户应当享有是否接受算法处理分析的自由，信息处理者和算法使用者应当设置用户随时退出免受算法分析的选项，并对以往的数据进行删除。算法是人类意志在技术层面的延伸，人们有理由选择拒绝使用此工具，退出算法的辐射范围，回归不被数字化的生活。（4）保证“数据透明”。模型的性能受制于原始数据，而对原始数据的每一项处理——收集、筛选、特征选择等均对算法歧视的发生有潜在的影响。因此不管出于算法审计或者事后追责的角度都需要确保训练数据具有可追溯性。数据可追溯性具体包括数据来源可查、数据内容可查、数据处理可查三个方面，因而就要求责任主体记录包括不限于数据获取的渠道、数量、时间等与数据的来源相关的事项；对数据内容进行记载及注释说明，并进行必要的备份；对训练数据具体处理步骤及目的予以记录，保证数据处理可查，且设置数据处理记录的访问权限，并保证记录不能被篡改。就后者而言，可引入区块链技术，将获取数据、处理数据等痕迹按照时间顺序以链条的方式组合成特定数据结构，通过密码学方式实现记录的不可篡改和不可伪造。

前治理的处理方式简单易于理解，但是前治理是要求企业放弃自己的一定的数据主权，势必引起企业的抵触；“数据被遗忘权”实践效果恐难达预期，一方面人们大概率不会频繁地行使这项权利，另一方面，为解决数据漂移问题，往往使用最新的数据用于模型的迭代，

“数据被遗忘权”或有利于个人隐私保护，但对于算法歧视的影响较为有限；更重要一点是机器学习之所以有现在的成就是因为海量数据的加持，从计算机科学的角度，数据量越大维度越高，机器学习模型性能才能取得更大的提升，对于前治理方案实际与计算机科学对于机器学习的期望有所违背，对于数据量极为有限或者数据维度被严格约束的样本，很难想象其可以训练出足够智能的模型，甚至可能适得其反，造成更严重的算法歧视。

4.2.2 事中治理：以算法治理为中心

算法设计的风险预防是消弭算法不合理影响，建立“规训的算法”，克服算法的技术偏私，引导科技向善。算法技术本身无意志，既要求开发人员应自觉防止技术异化又要通过法律、行业规范约束开发人员的行为。事中治理的主要内容包括：审查评估制度、道德内化与“过滤”函数。（1）健全算法审查评估制度。《个人信息保护法》第 54 条和 64 条规定了算法审查制度，条文中规定设计者具有自我审查与义务，必要情形下还需要监管机构和第三方机构的介入。后者弥补设计者和使用者对公平价值的重视不足，第三方机构应当是在该领域算法中有深入研究、具有中立性的权威机构，这有利于审查者和被审查之间的沟通、保证审查结果公正合理。审查内容方面，算法的审查过程中应当侧重于对算法设计基本原理、基本代码构建的深层次审查，审查标准主要是机器决策结果正当与否，尤其是决策结果出现差别待遇时，应对系统的运行逻辑进行合理性审查，对模型中使用的条件审查，确定是否适合且不超出必要范围，审查结果应当可留存化，建立专门的审查备案制度；审查时间方面，在模型上线使用前及运行过程中均有必要进行审查，由于受商业秘密的限制，算法设计或使用者只在必要的范围内公开算法技术，在运行过程中，需进行周期性审查，对于算法规制的日常化监管，可采取定期审查与不定期审查交叉的形式，监测算法运行情况，算法使用者对定期检查及公布检查报告负主体责任，对算法的可解释性负责[14]，即对决策过程、结果和意义等向社会公众及监管部门做出必要的解释。《个人信息保护法》第 55 条规定了风险评估制度，这成为算法使用者进行必要风险评估的法律依据。算法评估制度的构建和完善应关注：评估内容和标准的场景化。应注重场景的差异化，建立评估等级制度；注重技术性标准。法律上的风险评估标准应吸收技术性标准，注重技术性标准向法律标准的过渡。例如 2018 年电子技术标准化研究院发布的《人工智能深度学习算法评估规范》中，从代码实现的准确性、训练数据集的影响、算法功能实现的正确性、对抗性样本的影响等七个方面构建了算法可靠性评估指标体系。（2）算法设计中“道德内化”与添加“过滤”函数[15]。道德内化与“过滤”函数均是采用技术手段对算法进行约束。道德内化采用归纳偏置，归纳偏置主要是对解空间或样本产生过程的假设。常用的归纳偏置包括先验分布、使用正则项、网络结构特殊设计等。道德内化可以在模型训练中更改算法的目标函数限制歧视现象的发生。在设定归纳偏置的情形下，通过多个性能衡量指标确定最优模型，道德内化的方法就是借助归纳偏置寻找到一个性能最好且可以规避解空间存在歧视的模型。道德内化易实现易理解的方法是添加正则化项对算法的结构进行限制，该正则化项称为“道德化项”。目标函数值最小化是模型优化的目标，在添加道德化项之后，目标函数即为距离判别函数与道德化项之和。其中距离判

别函数是预测输出值与真实结果之间的距离，道德化项用于衡量算法歧视的程度、限制算法的结构，例如参数的复杂度等。两者呈负相关的关系，使得模型在寻找最优解时须平衡算法的准确性与公平性。数据是算法歧视产生的根源之一，数据是个人社会生活的数字化表达，数据中存在人们的价值预设，必然也存在着人类的偏见。这些偏见会在模型训练过程中以某些模式或规则的形式被学习下来，称之为算法偏见。最终是否形成歧视，取决于对归纳偏置的选择。在模型选择过程中，道德内化方案利用改变目标函数等归纳偏置方式避免算法偏见转变为算法歧视，以得到高准确率且兼顾公平的算法，该方案体现的思想是找到一种“正确的信息加工活动”。“过滤”函数同道德内化方案一样属“技治”的范畴，“过滤”函数应用于模型部署之前的测试阶段。旨在对“最优”模型的输出继续优化、消除歧视，主要方法是构造一个可以过滤具有歧视性特征的结果、或对有歧视性的结果进行修正的函数，将上一阶段的输出结果作为本函数的输入数据集，对该数据集进行二次变换，消除可能存在的歧视。“过滤”函数方法具有独立于模型、不会对模型的训练有任何干扰，其规制方案具有从属性存在的特点。

事前方案主要通过对源数据进行处理来避免算法歧视现象，但存在使用“正确”的数据并无法保证学习出“正确”的模型，事中治理方案采用了正则化或者修改学习表示的方式，自动自发地兼顾算法准确性与公平性，消除算法歧视的发生，一定程度上避免事前治理方案可能出现的“矫枉过正”情形。事中规制重心不在于数据中歧视的消除，而是用于训练的输入数据蕴含着某些偏见，亦可以得到公正的输出[16]。此外，事中治理方案有益透明原则的实现，事中治理方案并不需要具体的技术细节，审查工作可聚焦输出结果的公平性，可以有效消解算法的理解困难和数据公开危及个人隐私的弊端。

4.2.3 事后治理：以救济保护为中心

事中方案试图在歧视发生之前就将其消解，但并不能完全消除歧视侵权的可能，因此，事后的法律救济成为权利保障最后的利器，用以定罪惩罚与制止侵害。事后治理方案是在算法造成不合理的后果之后，使用法条来进行事后救济，通过向算法使用者或算法使用者追究相应责任来治理歧视。第三章对算法违法性认定已经进行了详细介绍。事后救济的基本思路：民事救济为主、行政刑事救济为辅。事后救济是规制算法损害的最后一道防线，如何最少地减轻损害后果，最大限度使被侵害人的权益恢复到圆满状态是对救济手段的考验。算法侵权根据不同场景不同程度会产生民事责任、行政责任乃至刑事责任。对于被侵权人的救济，应当从民事救济、行政救济和刑事救济三个维度出发，以民事救济为主，行政救济和刑事救济为辅，最大程度地保障算法相对人的数字权利。

在刑事救济方面，《互联网信息服务算法推荐管理规定》对刑法介入算法管理提供了法律正当性，但也应坚持刑法的谦抑性，除算法违反规定严重构成犯罪的，通过侵权赔偿和强制保险等民事手段及行政手段调整。在行政救济方面，监管部门作为国家权力机构，在日常监管中，应通过深入摸排企业算法应用的基本情况，全局把握风险，针对企业在应用算法进行决策中不缺位，形成算法干预机制，对结果进行纠偏，尤其对人工智能现阶段存在因为数

据使用不当、自身缺陷等原因造成自动化决策结果有偏，需要算法服务提供者对其进行有力的干预，充分保障用户权益。依职权对破坏市场行为、借助算法损害个体权益的行为依法作出具体行政行为，要求算法设计者及使用者整改并加以处罚，对造成消费者损失的，可责令算法设计者及使用者采取补救措施，包括不限于予以金钱赔偿等。行政主体监管算法需注重技术价值和法律价值的衡平，审慎行政问责。算法技术是人工智能的基础，在推动技术革新方面具有重要作用。算法问责应当坚持在行政执法和技术保护之间寻求良性平衡。在算法行政案件中，应当采取违法者主观过错标准。重视场景的差异化，完善问责标准，注重行为与责任相适应，应根据场景的不同设置不同的标准，构建科学的算法违法问责体系[17]。

算法在商业场景中的应用最为广泛、深刻。现有民事法律未对算法做出直接规定，但是算法的不当应用产生的后果本质上仍然是违约行为和侵权行为，其规制理论在于民法的诚实信用原则。虽无直接的法律规则援引，但民法中的诚实信用原则、意思自治原则等都能成为被侵权人寻求民事救济的法律依据。商业场景中借助机器决策，算法仅作为一类工具，有必要设立规制算法设计和使用者的法律规范行为准则。设计者应对其设计的算法负一定的担保责任，在实害发生时，应依据算法造成的损害就其应当履行而未履行或主观故意而导致实害发生的部分承担相应的责任，包括：数据可查义务、对算法的技术原理解释性义务及设计者的恶意操纵责任；算法使用者由于直接面对消费者，更可能直接侵害消费者的权益。因而应承担更为严格的义务。算法使用者对消费者负有忠实义务与注意义务[18]。算法使用者还负有注意义务、忠实义务，应防止对消费者造成损害、禁止实施有损终端客户利益的行为，不得以非法手段获取信息，不得滥用信息。此外，算法也会对消费者之外的个体产生影响。算法使用者掌握着大量的个人信息，尤其是个人敏感信息，负有对数据的存储、使用等的注意义务。由以上“可责”基础的构造，民事救济路径可梳理如下：

首先，注意算法缔约过程中的缔约过失和缔约完成后的违约责任。当算法使用主体采用算法对人们的数据进行分析时，应当做出事先的算法应用承诺。对于缔约时的承诺应当强化算法使用主体的责任，提高这一拒绝消费者反对权的格式条款的科学性，如果事先的承诺须知本身就有瑕疵，那后续的算法行为便缺乏正当性。此外，在算法应用过程中，应当严格按照缔约时的承诺进行，若出现过度分析等可能产生实质不合理后果的行为，可根据合同理论追究算法主体的民事责任，赋予相对人合同解除权，必要时可请求合理的赔偿。

其次，严格明确算法应用者的侵权责任。侵权责任的构成要求包括侵权主体、侵权行为、侵权后果、因果关系以及过错。算法的属性是一种代码产品，因此可参照侵权责任中的产品责任进行分析，算法开发者和应用者应为连带责任主体，在充分救济被侵害人的权利后，再进行内部责任划分。难以规制算法歧视的原因一方面在于算法侵权行为具有隐蔽性，另一方面在于责任主体具有复杂性，引入集体诉讼制度主要考虑是算法歧视的实害结果呈现规模化，单一受害者在权利、能力方面处于巨大劣势，两者之间的数字鸿沟失衡。集体诉讼成本更小，诉讼效率更高，更能保护受害者的合法权益，作为一种共同诉讼机制，受害人均有权利提起诉讼，且可以选举代表出席，在未出庭的情形下，可自主决定是否接受审判结果。

最后，当算法的不当应用给当事人造成实质不合理影响时，应当由当事人自由选择请求违约救济还是侵权救济。由于算法的隐蔽性和场景化特征，不同的民事纠纷场景可能在后果严重程度、纠纷复杂程度等方面迥然不同。因此，应当由当事人自己衡量举证责任和赔偿标准等问题，以最大化保护自身权利。在举证责任方面，我国《民法典》第 1034 条规定，有关隐私权的规定适用于个人信息中的隐私信息；没有规定的，适用有关个人信息保护的规定。《个人信息保护法》第 69 条规定，处理个人信息侵犯个人信息权益并造成损害的，个人信息处理者不能证明自己没有过错的，应当承担损害赔偿及其他侵权责任。前款规定的损害赔偿责任根据个人遭受的损失或个人信息处理者因此获得的利益确定；个人遭受的损失和个人信息处理者因此获得的利益难以确定的，应当根据实际情况确定赔偿金额。可以看出，我国的个人信息保护法将个人信息处理者的损害赔偿责任确定为推定过错责任。这一规定是基于这样一个事实：在数字时代，个人信息处理者与公众在掌握数字信息方面存在不平等，公众在证明信息处理人的过错时存在困难，且算法的应用者往往为大型公司甚至是政府，应采取过错推定原则，由应用者证明自身主观上有无过错，注意义务履行与否。若采用无过错原则有加重算法使用者之嫌。就责任方式，受害者向责任主体请求损害赔偿，设计者与使用者均承担全部责任，一方履行全部责任后，再进行内部责任划分，例如设计者是责任主体，算法使用者履行责任后有权向另一方追偿。若问题发现在产生实质损害过程中，设计者和使用者应分别在能力范围内积极、及时、完整地履行义务、采取补救措施，如暂停或停止使用算法、对已作出的决策不再执行等。否则，设计者和使用者在责任范围内承担相应的侵权责任，若一方已履行其应尽义务，侵权责任由另一方全部承担。

5 结语

本文介绍了算法歧视的相关概念，围绕算法歧视在机器学习过程中各阶段可能引起最终歧视发生的根源、算法歧视违法性认定以及算法治理的方法和分阶段治理等问题进行讨论分析，着重对算法歧视违法性判定方法进行梳理研究，主要原因是算法歧视具有很强的隐蔽性，往往存在认定困难，这也是算法歧视问题治理的一大难点，所以对此部分花较大笔墨展开，通过对实质平等和形式平等下的歧视、直接歧视及间接歧视的概念讨论，结合域内外对于不同歧视的判断方法，梳理出对于我国法律实践具有借鉴价值的判断方法；在本文核心内容即对算法歧视治理方面，从方法维度和时间维度进行梳理讨论，方法维度方面对法律治理和技术治理进行阐述，在机器学习不同阶段的算法歧视治理工作中，无不穿插着“技治”的方法，算法歧视的治理工作除了法律治理也需要更多技术专家的参与技术治理，进一步将技术治理的规范纳入法律制度以更具约束力，在工程技术实践中构建完备的机器决策系统全生命周期风险识别及防范能力，推进算法治理体系的完善。当然，在算法歧视的治理上，德治显然也具有重要作用，但德治具有非规范性的特征，因为量化困难，所以文中鲜少提及。时间维度方面，本文结合机器学习模型训练过程，将治理方案分为了事前、事中和事后三个阶段，进而结合技术治理和法律治理两种方法讨论不同阶段的治理方法：事前治理是一种事前的预防

性控制，主要内容是对数据与敏感属性的控制；事中治理是针对算法的治理，主要内容是构建算法审计、评估、备案三位一体的算法透明本土化方案和在算法中设置道德正则化项、过滤函数等约束条件以限制模型的技术治理方案；事后治理强调对实害结果启动救济措施，其中重点对民事救济涉及的若干问题进行了探讨。总而言之，本文认为对算法歧视的有效治理需要构建科学合理的算法歧视治理体系，而其中需要更多法律人员和技术人员的参与，将法治和技治方法充分结合，采用工程化的思想，对算法歧视产生不同阶段分而治之，构建事前风险预防、事中严格监管和事后多方救济的算法歧视治理框架，基于此逻辑，本文对其中内容进行有侧重的讨论。

[参考文献] (References)

[1] 张凌寒.算法权力的兴起、异化及法律规制[J].法商研究,2019,36(04):63-75.DOI:10.16390/j.cnki.issn1672-0393.2019.04.006.
[2] 陈林林,严书元.论个人信息保护立法中的平等原则[J].华东政法大学学报,2021,24(05):6-16.
[3] 崔靖梓.算法歧视挑战下平等权保护的危机与应对[J].法律科学(西北政法大学学报),2019,37(03):29-42.DOI:10.16290/j.cnki.1674-5205.2019.03.013.
[4] 王莹.算法侵害类型化研究与法律应对——以《个人信息保护法》为基点的算法规制扩展构想[J].法制与社会发展,2021,27(06):133-153.
[5] 胡元聪,冯一帆.大数据杀熟中消费者公平交易权保护探究[J].陕西师范大学学报(哲学社会科学版),2022,51(01):161-176.DOI:10.15983/j.cnki.sxss.2022.0112.
[6] 刘朝.算法歧视的表现、成因与治理策略[J].人民论坛,2022(02):64-68.
[7] 石颖.歧视的法律判断标准[J].河南大学学报(社会科学版),2022,62(01):42-46+153.DOI:10.15991/j.cnki.411028.2022.01.006.
[8] 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[J].比较法研究,2019(04):111-122.
[9] 沈伟伟.算法透明原则的迷思——算法规制理论的批判[J].环球法律评论,2019,41(06):20-39.
[10] 李晓辉.算法商业秘密与算法正义[J].比较法研究,2021(03):105-121.
[11] 魏远山.算法透明的迷失与回归：功能定位与实现路径[J].北方法学,2021,15(01):151-160.DOI:10.13893/j.cnki.bffx.2021.01.014.
[12] 张吉豫.构建多元共治的算法治理体系[J].法律科学(西北政法大学学报),2022,40(01):115-123.DOI:10.16290/j.cnki.1674-5205.2022.01.007.
[13] 刘友华.算法偏见及其规制路径研究[J].法学杂志,2019,40(06):55-66.DOI:10.16092/j.cnki.1001-618x.2019.06.006.
[14] 张欣.算法解释权与算法治理路径研究[J].中外法学,2019,31(06):1425-1445.
[15] 詹好.大数据时代下数据挖掘中的算法歧视研究[D].湖南师范大学,2020.DOI:10.27137/d.cnki.ghusu.2020.002312.
[16] 孟令宇.从算法偏见到算法歧视：算法歧视的责任问题探究[J].东北大学学报(社会科学版),2022,24(01):1-9.DOI:10.15936/j.cnki.1008-3758.2022.01.001.
[17] 肖冬梅.“后真相”背后的算法权力及其公法规制路径[J].行政法学研究,2020(04):3-17.
[18] 王莹.算法侵害责任框架刍议[J].中国法学,2022(03):165-184.DOI:10.14111/j.cnki.zgfx.20220511.009.