# 算法自动决策中人为歧视的法律规制

胡晓宇

沈阳工业大学，辽宁沈阳，110870；

**摘要：**随着大数据、人工智能等技术的大规模使用，算法自动决策已经广泛应用于我们的工作、生活和社会交往等领域当中。但是，算法并不是没有任何价值观倾向的技术工具，它本身所具有的“以人为基、人为而定”的特性决定了它很有可能会将人类既有社会的歧视偏见内化到自身之中，并不断对其复制放大，从而形成算法自动决策过程中的“人为歧视”，而且相较于一般意义的人为歧视来说，算法歧视具有更深藏不露的特点。算法歧视特有的技术黑箱性、技术复杂性和消解人化等特点使得它能够绕过传统反歧视法理体系的规制。从法理学的角度来看，围绕算法歧视问题可以做一些思考和讨论。因此，本文试图运用法理学的基本理论方法来探究大数据算法自动决策所形成的算法歧视问题。一方面，说明大数据算法歧视的本质不在于机器本身的自动意识，而是大数据算法下人类社会结构性偏见的一种代入体现，其根本原因就在于“人”。另一方面，在传统反歧视法理中，本文尝试对算法歧视所带来的“歧视意图”的认定、算法歧视对于“差别影响”归责的影响以及程序正义和个体权利的保障这三个层面展开思考，最后结合法理学方面的知识提出了应对算法歧视需要“多层次、系统化”的规制路径。

**关键词：**算法自动决策；人为歧视；法理规制；技术正义；算法透明度；责任分配

**DOI：**10.64216/3080-1486.26.03.060

## 1 算法歧视的本质：技术外衣下的人为偏见

对算法歧视进行有效规制的前提，是准确把握其本质。必须明确算法歧视并非技术自主的产物，而是人类社会既有偏见的数字化再现。其“人为”因素存在于算法生命周期的每一个环节。

### 1.1 数据投喂：历史偏见的“食粮”

算法的学习能力依赖于投喂给它的数据。然而，现实中数据并非理想的“事实群落”，而是带有社会偏见。比如公司历史上用以培训招聘算法的应聘人员信息显示，其工程师岗位一直由男性担任，则算法容易“学会”男性能代表优秀工程师的判断，然后这将是招聘时再分配给男性候选人更多的权重。但算法没有创造任何新的歧视，只是机械甚至不知情地拷贝了其中男性占优的不实数据，并将原有的偏见带入到了新的决策之中。数据变成了事实传递渠道，即偏见成为了占优数据，算法就是这一切成立后的证明者或者印证者。由“垃圾进，垃圾出”（GIGO）效应导致的歧视的根因并不来自算法模型，而是导致算法输入的这个社会环境本身所存在的问题，以及该社会进程中产生的数据信息不公平所致。

### 1.2 模型设计：价值偏好的“编码”

算法模型不是纯粹的数学公式，而是蕴含着设计者价值的选择与衡量结果。特征选择本身就是一个价值判断的过程。设计者要选择一些变量(例如年龄、性别、居住地、信用记录等等)来作为预测因子放进模型中，而不用另外一些变量。如果是用其他变量作为替代因素(比如用邮政编码代替种族),虽然不在字面上采用受保护的敏感属性，但却有可能达到精准歧视的效果。其次，目标函数的选择会影响整个算法的优化方向。如果目标函数是“利润最大化”，那么以这样一个目的来设计的信贷审批算法就会选择有资质却没有积累足够多的历史信用记录的低收入人群有风险有收益，而系统性地把这类人拒之门外。在这个例子中看似“理性”的选择实际上是加剧了社会阶层固化，因为我们的初始假设是整体上低收入人群的违约率要比高收入人群高。因此，算法模型的每一次参数调整、每一次权重分配，都是设计者价值偏好的编码过程。

### 1.3 反馈循环：歧视偏见的“自增强”

算法的输出结果又被以某种方式作为新的数据输入回算法本身，产生新的输出结果，如此往复构成一种螺旋式循环发展的过程。例如一个预测犯罪风险的算法模型如果依据于某一点位之前的犯案地点分布情况，以及社区治安情况等因素输出某区为高危区时，警察能够依据此结果勤加此区域巡查，并把遇到更多违规或者犯罪数据作为证明这一算法正确的基础证据。这样一来随着越来越多的新的犯罪数据反馈给算法并给与了佐证，下次再被判断为“高危”的机会就越大，这样一开始仅是零散存在的一些偏见会在长期以后逐渐积累放大成

一个“更大的错误”，“歧视性反馈循环”使最初的少量偏差无限放大，形成了霸权般的强制现实。算法由此变成了之前的被动者的角色，改变成了主动的再歧视机器，这种驱动因素依然还是人和社会以及算法三者之间相互作用的影响。

综上，算法歧视是在技术环境下人为因素汇聚而成的一种投射，在此过程中由一系列的人为因素聚集起来的数据与模型所触发，并依托于系统性、规模化的反馈机制而体现出算法歧视的客观表象特征。要让算法歧视被纳入到法律规制的视野中去、要认定相关的主体的责任，都离不开对这种“人为”的认识。

## 2 算法歧视对传统反歧视法理的冲击

算法人为歧视的隐蔽性、复杂性与去人格化特征，直接冲击了建立在人类行为基础上的传统反歧视法理框架，使其在应对新型歧视时面临解释力与适用性的双重困境。

### 2.1 对“歧视意图”认定标准的冲击

传统的反歧视法以及“差别对待”的法理均需要以决策主体具有“歧视意图”为前提，证明行为人的主观意图在于根据对象属于某种被保护群体，而对其进行不利对待，要想拿出令人信服的证据十分困难，尤其对于借助于算法做出的决策而言更是难上加难。在算法作出决定的过程中，即使相关开发者或者使用者称自己只是“算法中立”“结果源于数据”，不愿承担行为产生的后果，但平台的运行结果显然就是由其推动的，若想完全脱离决定结果去说明自己无“歧视意图”，对于平台而言极具挑战性。意图往往深藏于代码的繁杂性与数据的相关性当中，并披着“客观”的外衣，在面向算法这一法律无法释读的技术黑箱时，法律根本不具穿透的可能。因而采用“意图”作为规制基础的传统规制路径对于算法歧视已经难以奏效。

### 2.2 对“差别影响”归责模式的冲击

为应对“意图”证明的困难，法律发展出了“差别影响”（Disparate Impact）理论。该理论不要求证明歧视意图，只需证明某项看似中性的政策或实践，对受保护群体造成了不成比例的负面影响，且该政策与商业必要性等正当目标之间缺乏实质关联。然而，这一理论在应用于算法时同样面临挑战。

首先，随着算法的“商业必要性”或者“统计精确性”变得更加强大，算法开发者可以轻而易举地提交大量关于准确率、效率或成本节约等方面的统计、模型数据指标来论证自己算法具有“正当目的”，从而使人们信服其作出特定决定是有充分理由的。对于法院或者监管机构来说，在见识广泛以及专业知识有限的现状下，如果遇到非常专业的算法模型，通常会更加青睐算法提供商而不是其审查对象，难免会陷入到“技术权威”的迷雾之中。

其次，因果关系难以厘清。以外对于职场中女性歧视的因果关系非常容易证明，但在算法场景下，原本所需要做出的一个决定，可能需要调用成百上千个因素，运用一些复杂的非线性函数来最终计算出结果。这个时候，想要弄清楚到底哪个具体因素使得本来无辜的女性受到的是歧视待遇、到底是哪一个因素的参数设置有问题？还是说是数据偏差？就很难在技术上准确定义。而由于“罪魁祸首”难以明了，也让相关的责任难以下达。

### 2.3 对程序正义与个体权利保障的冲击

程序正义要求决策过程公开、透明，并且当事人具有申辩、救济的权利。但算法决策尤其是深度学习这样的“黑箱”模型都会让程序正义原则遭到极大破坏。比如某人因为评分低而得不到贷款，而机器判定时并不告诉你怎样就满足评分的要求，只有最简单的“综合评分不足”作为最终原因，由此导致无法进行申辩或者改正，更不要说质询该决策是否公正，这是一种程序上的无理性，也是侵害了个人的尊严与自主性的“算法暴政”，对于这种“解释的缺席”我们传统的程序性保障机能并不能起到很好的作用。

## 3 算法自动决策中人为歧视的法理规制路径

面对上述挑战，法律规制体系必须进行范式革新，从单纯的事后追责转向事前预防、事中控制与事后救济相结合的全链条治理。本文主张构建一个以“技术正义”为价值内核，以透明度、风险评估和责任分配为核心支柱的法理规制框架。

### 3.1 确立“技术正义”的核心价值理念

所谓“技术正义”是指技术发展应用必须要服务于社会的整体公平、正义和福利。对算法而言，“技术正义”的意义在于既要保证算法运行的高效还要体现价值，即在算法设计和应用时不能只顾及到技术上最优的结果，而且要使算法系统中包含着诸如公平性、无歧视以及可问责等伦理或法理的价值，将“公平性”作为我国法律、法规中的算法应用的基本原则。在立法和司法层面上，都应要求将“公平性”作为规则或者判例；比如在高风险领域应用的算法，应当在开发阶段就进行“公平性影响评估”，并将此作为市场准入的先决条件。树立“技术正义”观是对整个规制架构的价值引领以及合

法性支撑。

### 3.2 构建算法透明度与可解释性制度的基础原则

透明度是规制算法歧视的基石。没有透明，监督就无从谈起。透明度要求应分层分类：对于公共部门使用的、或对公民权利有重大影响的高风险算法，应适用最高的透明度标准，要求其公开基本原理、数据来源、主要特征变量及潜在的偏见风险。对于商业领域的算法，则可在保护商业秘密的前提下，要求向监管机构提供必要信息，并向受决策影响的个体提供“有意义的信息”。

与透明度紧密相连的是可解释性。法律应要求算法决策者，特别是当决策对个体产生不利影响时，必须提供“有意义的解释”。这种解释不应是技术术语的堆砌，而应是普通人能够理解的、说明决策关键依据的陈述。欧盟《通用数据保护条例》已对此做出了初步探索。未来，应通过立法或司法解释，进一步明确“有意义解释”的标准，推动可解释人工智能技术的发展，将法律要求转化为技术实现。

### 3.3 引入算法影响评估与审计机制

借鉴环境影响评价的有效经验，构建“算法影响评估”制度，要求算法的设计者、部署者在上线前以及在线过程中，对可能产生的社会负面效应、特殊人群的影响等尤其是歧视问题进行系统的评估，并将评估报告公之于众。同时，设立独立第三方“算法审计”，即由具备一定技术和法律知识的独立机构，对算法系统是否公平、公开、合法等进行定期或者不定期检查，审计内容涉及到数据质量、模型设计、决策结果等方面，算法审计不合格的要被要求整改、暂停使用或下架，用制度化的手段撬动外部监督力量干预破解“技术黑箱”。

### 3.4 明确多元化的责任分配链条

规制算法歧视，最终要落实到责任主体上。必须打破“算法免责”的幻想，构建一个覆盖算法设计、开发、部署和使用全链条的多元化责任体系。

首先要认清算法设计者、开发者的责任。他们是算法的“创造者”，对于算法的基本结构以及存在的隐患有着比他人更为清楚的认识，应对其施加一定的注意义务，当存在明显的缺陷或者没有尽到最基本预防偏见发生的义务而造成歧视的结果，则应该由其承担相应责任。

其次要落实算法部署者与使用者的责任。部署和使用算法的人是算法作出决策的直接受益者和风险控制人。部署者应该成为最终担责的人，他们不能因为“技术外包”就可以摆脱责任，也就要对此负责。这就要求他们除了需要对自身使用的算法进行尽职调查以外，还要进行不断的跟踪检验，并建立好相应的内部申诉、救济机制。

最后不能忽视监管者的责任。监管部门要由被动的审批者转变为积极主动的引领者、监督者。一方面要提出明晰的技术标准和伦理要求，督导算法审计及影响评估工作的落实；另一方面也要积极建设受侵害公民获赔便捷的救济渠道。

基于“设计者-部署者-监管者”三位一体的责任链条来构建算法责任体系，在算法全生命周期的每一个环节都落实到人，用制度的牢笼将各类人为歧视套牢并予以震慑和矫正。

## 4 结语

算法自动决策中的人为歧视，是数字时代对法治文明提出的一项深刻考验。它以技术之名，行歧视之实，挑战着法律对公平正义的守护。本文的分析表明，应对这一挑战，我们必须超越对技术表象的恐惧或迷恋，直击其“人为”的本质。算法歧视并非不可逾越的技术鸿沟，而是人类社会既有偏见在数字空间的延伸与固化。为此，法理规制体系必须进行一场深刻的自我革新。我们需要从“技术正义”的价值高度出发，重塑法律对技术的立场

最终，对算法人为歧视的规制，不仅是法律制度的完善，更是一场关于我们希望建立何种数字未来的社会抉择。它要求我们以法律之理性，驾驭技术之力量，确保在迈向智能社会的征程中，不让任何一个人被数据和代码所定义、所抛弃，真正实现技术进步与人类福祉的同频共振。

## 参考文献

[1]钟晓雯.算法推荐网络服务提供者的权力异化及法律规制[J].中国海商法研 究,2022,33(04):63-72.

[2]曾雄,梁正,张辉.欧美算法治理实践的新发展与我国算法综合治理框架的构建[J].电 子政务,2022,No.235(07):67-75.

[3][美]托马斯·索威尔.歧视与不平等[M],刘军译.北京:中信出版集团股份有限公司,2021.

作者简介：胡晓宇（2000—），男，汉族，山东省威海市乳山市，沈阳工业大学/研究生，法理学方向。