·科技与社会·

人工智能算法歧视及其治理

汪怀君,汝绪华

(中国石油大学(华东) 马克思主义学院,山东 青岛 266580)

摘　要:人工智能的快速崛起,在人类事务领域引发无法想象的革命的同时,愈益严重的算法歧视现象也引发了人们普遍的担忧与焦虑。AI 算法绝不仅仅是一项单纯的技术性问题,其蕴含的政治风险、安全风险与伦理道德风险同样不容忽视。实际上,算法歧视不仅会导致种族歧视、性别歧视等严重社会后果,在一些关键领域或特定情境下,还会侵害公民权利、自由,甚至危害其生命安全。因此,对算法歧视善加治理,才能为算法及其行业的健康发展创造良好的环境。

关键词:人工智能;算法歧视;算法黑箱;伦理

中图分类号:G301　　**文献标识码**:A　　**文章编号**:1674 - 7062(2020)02 - 0101 - 06

人工智能(Artificial Intelligence,简称 AI)一词由美国人约翰·麦肯锡(John McCarthy)在 1956 年提出,指的是依托计算机运用数学算法模仿人类的分析、推理和思维能力。人工智能发展离不开算法,算法是人工智能的灵魂,算法的优劣直接决定了人工智能水平的高低。所谓人工智能算法(AI Algorithm)指的是:"在计算机科学中用于描述一种有限、确定性和有效的问题解决方法,适合作为计算机程序来实现。"[1]随着人工智能的快速发展,在交通运输、家务劳动、医疗保健、娱乐产业、雇佣与工作管理、公共安全、低能耗社区和教育八大社会领域,算法正在逐步改变我们的日常生活。然而,与人们对依赖大数据、机器学习的 AI 算法会带来更加客观公正判断的殷切期望形成巨大落差的是:算法绝非中立、公正的,随着算法全面渗透生活世界,算法歧视现象高发,已不容漠视。

一　算法歧视的概念与主要形态

在当今人工智能高速发展的关键时刻,频发的算法歧视引起的巨大争议正成为阻碍人工智能发展的痛点与顽疾。美国白宫 2014 年、2016 年发布的大数据研究报告都关注了算法歧视现象。2015 年 11 月,欧盟数据保护委员会(EDPB)发布的《应对大数据挑战》(*Meeting the challenges of big data*)同样强调了大数据和算法中的歧视问题。2018 年 11 月,皮尤研究中心发布的《公众对计算机算法的态度》(*Public Attitudes Toward Computer Algorithms*)调查报告也显示:58% 美国受访者认为计算机程序将始终反映出一定程度的人为偏见。算法歧视(Algorithmic Bias)指的是人工智能算法在收集、分类、生成和解释数据时产生的与人类相同的偏见与歧视,主要表现为年龄歧视、性别歧视、消费歧视、就业歧视、种族歧视、弱势群体歧视等现象。AI 算法应用正变得越来越普遍,越来越多的利益分配和大数据直接相关,尤其是在算法决策应用日益广泛的教育、就业、福利补贴发放、刑事司法、公共安全等重要与高价值领域,算法歧视可能会导致严重的政治与道德风险。

【收稿日期】 2019 - 03 - 26

【基金项目】 国家社会科学基金项目"生态女性主义视阈下女性符号消费的伦理研究"(16BZX107);中央高校基本科研业务费专项资金资助项目"马克思主义'人与自然双重解放'思想研究"(19CX04032B)

【作者简介】 汪怀君(1978 -),女,山东临清人,中国石油大学(华东)马克思主义学院副教授,研究方向为科技伦理;
汝绪华(1976 -),男,山东济南人,中国石油大学(华东)马克思主义学院副教授,研究方向为算法治理。

第一，算法引起的种族歧视更隐蔽。有形的种族歧视容易精准打击，无形的种族歧视却难以防范。被嵌入种族歧视代码的算法中隐藏的“歧视特洛伊木马”在人工智能“客观、公正、科学”的高科技包装下更容易大行其道，在算法黑箱的遮掩下更隐蔽。以人脸识别技术为例，在这个人脸识别的年代，人脸识别技术的发展与广泛应用，再次将种族歧视问题摆上了台面。麻省理工学院的一项研究表明：“当使用各种人脸识别算法来识别性别时，算法将肤色较深的女性误分类为男性的比例为34.7%；而对肤色较浅的女性的分类最大错误率不到1%。”[2]随着人脸识别系统变得标准化，并逐步应用于学校、体育馆、机场、交通枢纽特别是警务系统，人脸识别技术的种族歧视引发的对有色人种群体的新型伤害愈发突显。更需要警醒的是基于减少偏见而应用的算法反而加剧了种族歧视，PredPol已经在美国的几个州使用，它是一种旨在预测犯罪发生时间和地点的算法，目的是帮助减少警察中的人为偏见。但在2016年，当人权数据分析小组将PredPol算法模拟应用于加利福尼亚州奥克兰的毒品犯罪时，它反复派遣警务人员到少数族裔占高比例的地区，无论这些地区的真实犯罪率如何。近年来，算法引发的种族歧视现象层出不穷，充分说明了虚拟世界反种族歧视的紧迫性与重要性。

第二，算法造成的性别歧视实质上是现实世界长期存在的性别歧视观念在虚拟世界中的延伸。大数据是社会的产物，人类不自觉的性别歧视会影响对大数据进行分析的AI算法，可能无意中强化了就业招聘、大学录取等领域中的性别歧视。“算法的核心是模仿人类决策……换句话说，算法不是中立的。”[3]比如，当用工单位在自动简历筛选软件中输入“程序员”时，搜索结果会优先显示来自男性求职者的简历——因为“程序员”这个词与男性的关联比女性更密切；当搜索目标为“前台”时，女性求职者的简历则会被优先显示出来。当“谷歌翻译”将西班牙语的新闻文章翻译成英文时，提及女性的短语经常会变成“他说”或者“他写”。2019年11月，戴维·海涅迈尔·汉森(David Heinemeier Hansson)质疑为什么苹果信用卡(Apple Card)给他的信用额度是他妻子的20倍，他的妻子实际上拥有比他更好的信用评分。尽管性别歧视人人喊打，但它仍不时癔症般作祟。

第三，算法引致的年龄歧视是工作场合歧视中最难以证明的一种形式。在就业招聘、员工管理中，就业者的姓名、性格、兴趣、情感、年龄乃至肤色等数据，往往悉数被采集，运用大数据算法非常方便对年龄等数据进行筛选与评估。比如，对于寻求新工作的不同年龄段的人来说，他或她的日常工作可能包括搜索互联网工作网站和提交在线申请。从表面上看，这似乎是一个非常透明和客观的过程，将所有申请人置于一个只有经验和资格的公平竞争环境中，但实际上年龄歧视无处不在。2016年，ACCESSWIRE的ResumeterPro项目组发现，在人工有机会审查之前，高达72%的简历会被申请人跟踪系统拒绝。这是通过复杂的算法来完成的，可能会导致基于不准确假设的无意识歧视，雇主则可以使用这些算法根据年龄专门丢弃申请。令人不安的是，这种公然的歧视很难被发现，因为申请人很难证明拒绝是由于年龄原因造成的。比利亚雷亚尔诉雷诺烟草公司案(Villarreal v. R. J. Reynolds Tobacco Co.)提供了一个教科书示例，说明算法如何导致隐藏的年龄歧视。比利亚雷亚尔曾多次在网上申请雷诺烟草公司工作，但一直未收到回复，直至“该公司根据年龄筛选在线申请人、但未向任何被拒绝的候选人透露此事”的丑闻被举报。

第四，算法“算计”的消费歧视令人防不胜防。算法时代，商业互联网平台通过深挖消费者以往的消费数据与浏览记录，为消费者进行精准数字画像与数字建档，让算法洞悉消费者偏好，可以轻松地针对不同地域、不同时段的用户差别定价，以实现利润最大化。针对不同的细分市场，同样的产品与服务是可以通过差别定价来获取更多的利润的。作为一种正常的商业策略，差别定价是企业追求利润最大化的合理定价行为，在机票、酒店、电影、电商、出行等价格易有波动的领域都存在差别定价。只要企业定价行为公开透明，消费者也愿意接受，这样的行为就不存在欺诈。但差别定价又有一个明确的边界：商家不能针对某个具体的个人或特定群体歧视性提价。2017年，非营利组织ProPublica通过对加利福尼亚州、伊利诺伊州、德克萨斯州和密苏里州的保险费和支出的分析表明，一些主要保险公司向少数族裔社区收取的费用比其他具有类似事故费用的地区高出30%。亚马逊的购物推荐系统、在线旅游网站奥比兹(Orbitz)、携程、滴滴打车等都曾涉嫌利用大数据杀熟，实行差别定价，对老用户进行价格歧视。

第五，算法对弱势群体的歧视无所不在。算法模型设计上存在的偏见会造成弱势群体在雇佣评估、信贷、住房、保险甚至刑事司法上遭遇歧视。美

国联邦贸易委员会(FTC)的调查发现:广告商倾向于针对生活在低收入社区的群体推送高息贷款信息。样本不平衡、有意遗漏或倾向性选择,同样会造成弱势群体歧视。任何有着浓重或者不常见的口音的人都可能有过被 Siri 或 Alexa 误解的经历,如果某一种特定的口音或是方言没有足够的样本数据,这些语音识别系统就很难听懂他们究竟在说什么。自然语言科技支撑着与顾客的自动交流,也被用来挖掘网络和社交媒体上的公众意见和梳理文本材料里的有效信息,这意味着基于自然语言系统的服务和产品有可能歧视特定族群。在就业招聘领域,情况可能更糟,“雇主们正在转向以数学方式筛选工作申请的方式,即使是错误的,他们的判决似乎也无可争议——他们倾向于惩罚穷人”[4]。招聘公司 HireVue 在全球拥有 700 多家客户,该公司通过在线“视频面试”或“基于游戏的挑战”来评估求职者的工作资格,但许多年长的员工都不玩电子游戏,玩游戏的女性也比男性少,因此,该公司的专用算法对年长员工、女性等群体造成了就业歧视。貌似客观中立的算法对弱势群体的“算计”更隐蔽,危害更甚。

二　算法歧视滋生的肇因

AI 算法时代,随着人工智能技术的广泛应用,尤其是在决策、刑事司法等高价值领域的运用,如何确保算法遵循人类的伦理与道德变得越来越重要。实际上,算法歧视问题比很多人意识到的要严重得多,在一些关键领域或特定情境下,算法歧视不仅会侵害公民的权利、自由,甚至危害其生命安全。因此,厘清算法歧视滋生的肇因,既是 AI 算法行业健康发展的重要保障,也是预防与治理算法歧视的必要手段。

*第一,算法研发者的偏见。*毋庸讳言,算法歧视很大程度上是社会上早已存在的种种偏见、歧视与刻板印象在虚拟世界的投射。算法及其决策程序终究是由人设计的人造物,研发者的利益与价值取向难免会嵌入其中。正如麻省理工媒体实验室专家拉胡尔·巴尔加瓦所言:“算法没有偏见,我们有。”[5]由于算法黑箱的“阻隔”,普通人看到的只是结果而非决策过程,很多人在毫不知情的情况下承受着种种隐形的歧视与精准的靶向不公正。2016 年,脸书的“热门话题榜(Trending Topics)”陷入新闻偏见门,其管理者有意识地通过“注入工具”引导用户去浏览他们认为重要的话题,不合乎管理者喜好的话题还会被拉黑。算法偏见造成的最棘手的歧视症结还在于:一方面,即使算法研发者主观上没有性别歧视、种族歧视等观念,通常也难以完全屏蔽刻板印象与偏见。原因在于:研发者的技术思维与逻辑,使其易陷入“只见数字不见人”的唯技术主义陷阱;同时,由于缺乏对所设计算法模型适用领域相关知识背景和价值规范的深入洞察,对数据承载的描述信息缺乏深刻了解,往往难以把该领域的全景与细节用恰当与精准的代码表达出来。另一方面,人类歧视尚能通过自觉或外在压力进行纠正,但算法却可以通过数据挖掘、关联分析发掘呈现隐藏于大数据中的歧视与偏见;非但如此,技术中立的科学外衣、无法识别的算法黑箱操作等都极大增加了算法歧视解决的难度。

*第二,样本与训练数据的偏见。*样本与训练数据通常被喻为算法的“教科书”,机器学习的效果与样本平衡与否、训练数据多少密切相关。训练样本不能太少,各个类别的样本数量差别不能太大,太少、差异大的数据无法有效代表数据的整体分布情况,容易造成过拟合。比如,用于图片分类的深度神经网络通常采用的 ImageNet 数据集就存在样本不平衡与训练数据缺陷,“超过 45% 的 ImageNet 数据来自美国……中国和印度加起来只贡献了 ImageNet 3% 的数据……这部分解释了为什么计算机视觉算法会给一张传统美国新娘的照片贴上‘新娘’‘礼服’‘女人’等标签,而一张北印度新娘的照片只有‘表演艺术’和‘礼服’两个标签”[6]。因为机器学习算法需要大量数据才能完美理解。数据的假阳性和假阴性两类误差也可能误导算法判断,“在一般的数据统计中,数据的假阴性与假阳性可能无足轻重,但是在司法程序中,即使是再轻微的数据偏差,都有可能造成事实认定错误及司法不公,尤其是‘假阳性’错误可能将无罪之人认定为有罪”[7]。非营利组织 ProPublica 对犯罪风险评估算法(COMPAS)研究后发现,黑人更倾向于被错误评估为高犯罪风险,其概率大约是白人的两倍。这种由算法黑箱运作的不公正的风险评估严重影响了司法公正。

*第三,算法研发公司以及购买企业的利益诉求。*算法研发的最终目的还是为了应用牟利,因此,研发公司与购买公司的意图无疑会深刻地反映在算法设计中。如果某些企业把利润追求凌驾于企业社会责任与社会公德之上,算法歧视也就难以避免。在 2016 年美国总统大选中,脸书的排名算法疯狂推荐假新闻,美国新闻聚合网站 BuzzFeed 的调查发现,在美国大选的这几个关键月份,来自恶作剧网站和

超党派博客的20个表现最佳的虚假选举故事在脸书上产生了871.1万次的分享与评论。脸书却获得了不菲的红利,2016年第四季度财报显示,其营收高达88.1亿美元,高于分析师们预测的85亿美元。算法研发公司侵犯公众隐私,同样会造成算法歧视。比如,滴滴司机在接单前,不仅可以看到乘客过往顺风车的详细记录,还可以看到一些司机对该乘客的露骨评价,如"肤白貌美""安静的美少女"等,对公众隐私的侵犯、性别歧视使得滴滴打车备受诟病。

第四,算法自身原因造成的歧视。主要表现为四方面:一是,算法黑箱,自我解释性差。AI算法系统非常复杂,就像一个"黑箱",给出的只是一个冰冷的数字。它是如何得出结论,依据什么,无人知晓。某些情况下,甚至连设计它们的工程师都无能为力。二是,算法极其复杂,涉及大量专业知识,难以理解,即使专业人士很多时候都未必能在短时间内了然其设计结构,对于普通人来说,若想洞悉某一种算法的奥秘,难度可想而知。这是算法难以审计、难以监管的重要原因,也是算法歧视难以预防之所在。三是,目前,人工智能领域陷入了概率关联的困境:不问因果只关问相关性,只做归纳不做演绎。算法可以学会识别和利用这样的一个事实,即一个人的教育水平或家庭住址可能与其他的人口信息相关联,族裔歧视、区域歧视和其他偏见可能被它"理所当然"视作"事实"。四是,算法已经自我学会歧视,也容易被"教坏"。普林斯顿大学的艾琳·卡利斯坎(Aylin Caliskan)等学者使用内隐关联测试(IAT)量化人类偏见时发现,在利用网络上常见的人类语言进行训练时,机器学习程序从文本语料库中自动推导的语义中包含了类似人类的偏见[8]。同时,算法容易被人类"教坏",清华大学图书馆的机器人"小图"、微软的Tay聊天机器人都曾因被教坏而"下课"。

第五,西方中心主义意识形态也是算法歧视普遍存在的重要原因之一。世界上最强大的互联网公司以及最先进的算法技术大多集中于西方国家,因此,西方国家根深蒂固的西方中心主义意识形态难免以"内隐性社会认知"的方式嵌入算法之中。即使算法研发者力求公正、客观,但其深入骨髓的偏见与潜意识也会影响其设计以及样本、数据的选择。人类倾向于将固有价值注入规则,AI算法歧视只不过是已有文化与认知的延伸。2016年,一个名为罗萨莉娅(Rosalia)的学生发现了一个奇怪的现象:用谷歌搜索"看起来工作不专业的发型",结果图片绝大多数都是一头自然卷的黑色人种女性。相反,如果搜索"看起来工作专业的发型",结果是铺天盖地的白人女性。其实白种人的优越地位不止体现在发型的评价上,假如你在谷歌图片搜索"男人",你会发现绝大多数结果都是各种年龄段的白人男性;而如果你搜索"女人",你会发现绝大多数结果是年轻的白人女性。尽管西方国家高举种族平等大旗,但实际上,基于西方中心主义的无形的偏见与歧视却根深蒂固。

三　算法歧视的治理策略

面对来自四面八方的强烈批评与信任危机,越来越多的国际组织、国家或地方机关、行业协会、网络科技公司、学术机构以及非营利组织正纷纷加入到算法歧视治理的行列中来。唯有对算法歧视善加治理,才能为算法及其行业的健康成长创造良好的环境。

第一,建立与完善算法立法,加强对算法的制度化监管。现今,算法决策的影响越来越重大,特别是在政府决策、警察和刑事裁决等领域,算法决策意外风险越来越高,算法歧视防不胜防,因此,亟需建立与完善AI算法立法,加强对算法的制度化监管。在算法立法方面,美国的力度是最大的。2017年12月,纽约市议会通过了《算法问责法案》,这是第一部审查算法偏见与歧视的法案,该法案将试图为城市政府机构使用算法的方式提供透明度。2019年4月,美国两位民主党参议员布克(Cory Booker)和怀登(Ron Wyden)联合提出了《2019算法问责制法案》(*Algorithmic Accountability Act of* 2019),试图对人工智能机器学习中的偏见和个人敏感信息使用问题进行规制。美国国会两院正在讨论《人工智能未来法案》《人工智能就业法案》等多部着眼于确立美国未来人工智能领导地位的法案,也都非常注重防范算法歧视。2018年生效的欧盟《统一数据保护条例》(GDPR)不仅将个人敏感数据排除在人工智能的自动化决定之外,而且着重强调了算法可解释性、算法审计,力图解决算法歧视问题。另外,韩国、爱沙尼亚等国家也纷纷提出了人工智能法案。立法规制AI算法,治理与预防算法歧视,越来越多的国家在行动。

第二,建立算法从业人员以及AI行业的伦理规范,从源头遏制与预防算法歧视。预防算法歧视,建立算法研发与数据采集从业人员的伦理规范以及塑造他们的社会责任感非常重要。算法研发工程师、

样本数据采集员要遵守一些基本的伦理准则，包括客观公正、有益、健康、不作恶、包容性、多样性、透明性、责任性、保护隐私等，并将人类社会的法律、道德规范嵌入算法系统。防范算法歧视，也必须建立 AI 行业的伦理规范。2017 年，电气电子工程师协会（IEEE）发布了《人工智能设计的伦理准则（第 2 版）》（*Ethically Aligned Design Version* 2），涵盖一般原则、价值嵌入、经典伦理问题等 13 个方面，是全球范围内系统、全面阐述人工智能伦理问题的重要文件。美国公共政策委员会、英国人工智能委员会、德国数据伦理委员会等也纷纷提出了规范人工智能算法发展的伦理规范。遏制算法歧视，同样需要国际组织的努力。2015 年 11 月，《应对大数据挑战》报告提出了“公平能否算法化”的问题，特别指出：要警惕大数据对穷人或者弱势群体的歧视。2019 年 5 月，经济合作与发展组织（OECD）通过了首部人工智能的政府间政策指导方针，确保人工智能的系统设计符合公正、安全、公平和值得信赖的国际标准。治理算法歧视，专业学术组织的参与更是必不可少，2017 年 1 月，未来生命研究院召开主题为“有益的人工智能（Beneficial AI）”的阿西洛马会议，达成了 23 条人工智能原则。从源头上预防算法歧视刻不容缓，全球在行动。

第三，网络科技巨头纷纷进行算法伦理自律，把伦理嵌入算法设计与应用中，以预防算法歧视。道德所以有意义，乃是因为道德行为是具有自主性者之理性行为。2016 年，微软提出了人工智能研发的六项核心设计原则，着重强调了人工智能的透明、防范算法歧视、个人隐私保护、算法问责等内容。2018 年，谷歌发布了 AI 研究七大准则以指引未来公司的 AI 工作，这些原则不是理论概念，而是能够积极指导研究和产品开发、影响商业决策的具体标准。网络科技巨头纷纷认识到对算法进行伦理审视与约束的重要性与紧迫性。2016 年，亚马逊、谷歌等几家全球大公司还共同成立了 Partnership on AI，研究与制定 AI 技术的最佳实践。然而，由于人工智能的决策或行为的影响往往是分布式智能体（设计师、开发者、用户、硬件、软件）互动的结果，导致了责任分配也是分布式的。现有的伦理框架涉及个体的责任，却不能处理分布式的责任分配，这是英国女性乳腺癌筛查软件漏检事故中相关各方互相踢皮球的根本原因。“直到最近，新定义的伦理理论才考虑到分布式的智能体，它所提出的理论依赖于合约和侵权责任并采用了一个完美无缺的责任模型……它在设计者、监管者和用户之间分配伦理责任，模型在预防作恶和培养善心方面起着核心作用，因为它促使所有参与者采取负责任的行为。”[9]

第四，推进算法设计的优化与技术进步，以算法治理算法歧视。算法歧视可以通过技术的改进和创新予以克服。目前，越来越多的学者投入到算法设计优化与预防算法风险新技术的研究与开发中来。微软程序员亚当·凯莱（Adam Kalai）与波士顿大学的科学家合作研究一种名为“词向量”的技术，目的是瓦解算法中存在的性别歧视。加利福尼亚大学伯克利分校和马克斯普朗克信息学研究所合作提出了一种能够自我解释的算法，这种被称为“指向和对齐”的系统有助于让人类理解机器学习的决策过程。在推进人工智能算法因果推断的研究上，算法科学家已经形成共识，并不断加大研究投入，寻求突破。AI 审计也是一种备受业界推崇的预防算法歧视的好方法，通过使用机器学习本身系统性地检查原始的集群学习模型，以识别模型和训练数据中的偏见。令人欣喜的是，网络科技巨头纷纷加入算法优化与改进的行列中来，微软正构建可防止 AI 算法出现偏见的新工具，谷歌研发了“阈值分类器”，通过改进机器学习系统来避免歧视。脸书也发布了 Fairness Flow，该工具会自动警告某种算法是否根据检测目标的种族、性别或者年龄，做出了不公平的判断。网络科技巨头已经认识到算法歧视的严重性，正投入越来越多的资源加强研究与研发。

第五，破除西方中心主义意识形态，同时，提升公众的算法素养对抗算法歧视。除了算法发展的技术性局限之外，在西方国家，算法研发者、开发公司根深蒂固的西方中心主义意识形态深深嵌入算法是算法歧视与偏见滋生的重要诱因。“AI 只是客观地通过算法捕捉我们生活的现实世界，而我们的世界里充满了歧视与偏见，AI 就不可避免地被‘带坏’……人类消除了偏见与歧视，AI 才可能重新学好。”[10]众所周知，训练数据越多、越全，算法也就越准确。然而，最流行的数据库往往以西方社会与文化为核心，训练图片识别算法常用的开源数据库 ImageNet 和 Open Images、训练人脸识别工具常采用的源域数据集（Source domain dataset）等莫不如此，这是造成使用公共开源数据库训练的图片识别算法只能识别身着西方婚纱的新娘、却无法识别身着印度纱丽的新娘的根本原因。因此，要摒除算法歧视必须破除西方中心主义意识形态，树立公平正义、人类命运共同体理念。提升公众的算法素养也是对抗

算法歧视的非常有价值的方法,虽然让公众洞悉各种算法的工作原理是不可行的,但给其教导一种健康的“知情怀疑论”则是必要的。与此同时,更多披露由人工代理人辅助的决策和行动,把算法素养与算法决策的透明度相结合可能会非常有效。

【参 考 文 献】

[1]SEDGEWICK R, WAYNE K. Algorithms[M]. 4th ed. California: Addison - Wesley Professional, 2011:4.

[2]BUOLAMWINI J, GEBRU T. Gender shades: intersectional accuracy disparities in commercial gender classification[J]. Proceedings of machine learning research, 2018(81):1 - 15.

[3]MANN G, O'NEIL C. Hiring algorithms are not neutral [EB/OL]. (2016 - 12 - 09)[2019 - 03 - 18]. https://hbr.org/2016/12/hiring - algorithms - are - not - neutral.

[4]O'NEIL C. How algorithms rule our working lives[N]. The guardian, 2016 - 09 - 01.

[5]BHARGAVA R. The algorithms aren't biased, we are[EB/OL]. (2019 - 01 - 29)[2019 - 03 - 28]. https://www.kdnuggets.com/2019/01/algorithms - arent - biased - we - are.html.

[6]ZOU J, SCHIEBINGER L. AI can be sexist and racist: it's time to make it fair[J]. Nature, 2018(559):324 - 326.

[7]王燃.大数据侦查[M].北京:清华大学出版社,2012:67.

[8]CALISKAN A, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human - like biases[J]. Science, 2017, 356(6334):183 - 186.

[9]TADDEO M, FLORIDI L. How AI can be a force for good [J]. Science, 2018, 361(6404):751 - 752.

[10]HUTSON M. Even artificial intelligence can acquire biases against race and gender[EB/OL]. (2017 - 04 - 13)[2019 - 12 - 06]. https://www.sciencemag.org/news/2017/04/even - artificial - intelligence - can - acquire - biases - against - race - and - gender.

AI Algorithm Bias And Its Governance

WANG Huai - jun, RU Xu - hua

(*School of Marxism*, *China University of Petroleum*, *Qingdao Shandong* 266580, *China*)

Abstract: Even though the rise of the new generation of artificial intelligence brings surprises and higher productivity to human beings, its serious algorithmic bias has also caused people's general concerns and anxiety increasingly. The AI algorithm is not only a technical issue, but also a political and ethical issue. AI Algorithms continue to reshape human production and lifestyle. How to ensure that algorithms follow human ethics and morals is becoming increasingly important. In fact, algorithmic bias will not only lead to serious social consequences such as racial discrimination and gender discrimination, etc. It will also infringe civil rights, freedoms, and even endanger people's lives in key areas or specific situations. Therefore, the good governance of algorithmic bias will be able to create a good environment for the healthy development of the algorithm and its industry.

Key words: artificial intelligence; algorithmic bias; algorithm black box; ethics

(责任编辑 许玉俊)