

算法歧视与我们的距离

叶慧莹

福建师范大学 心理学院, 福建福州, 350100;

摘要: 歧视无处不在, 性别、种族、年龄、学历等话题常常引发社会热议。算法歧视, 指数据驱动的计算决策会再现先存的社会歧视模式, 从而加深不平等。本文探讨的算法歧视是客观存在的专业术语, 非仅主观道德判断。算法因其精确缜密的计算能力和较低廉的成本而备受青睐, 逐步融入各种常见领域并代替人类做出关键决策, 从医疗决策到金融投资, 乃至司法量刑, 算法能在一定程度上避免人类的主观性, 比人类决策更加理性、公正、无偏, 但事实上算法歧视已经渗透到各行各业, 群策群力纠偏是大众未来关注和努力的方向。本文就以算法歧视为研究对照, 梳理当前研究探索得到的分类和成因, 探讨新兴技术带来的法律挑战和解决方案。

关键词: 算法歧视; 偏见; 数字科技; 人工智能

DOI:10.69979/3029-2735.25.4.065

1 引言

中国互联网络信息中心 25 日在京发布第 54 次《中国互联网络发展状况统计报告》, 报告显示, 截至 2024 年 6 月, 我国网民规模近 11 亿人 (10.9967 亿人), 较 2023 年 12 月增长 742 万人, 互联网普及率达 78.0%。越来越多的行业依赖于算法来辅助工作, 生活中每一次刷脸支付、人脸认证乃至每一次网购所产生的数据都在记录着我们的数字足迹。数据在真实呈现生活的同时, 也越来越多地将人类社会的歧视投射到机器上, 这不仅会损害虚拟空间的秩序, 也会对现实社会的公平正义产生负面影响。在传统歧视事件中, 表现歧视的主体是人类, 大众倾向于道德谴责, 然而, 随着计算机技术的发展和应用, 算法也成了新的歧视主体。

加纳裔科学家乔伊发现, 她使用的人脸识别系统总是检测不了自己的脸, 除非她戴上一副白色面具。乔伊进而测试了许多商用软件, 发现白人男性面孔被识别的正确率远远高于其他人群, 她称之为“编码凝视”, 她的团队因此创立了国际组织——“算法正义联盟”, 对算法歧视的声讨和反思也日益受到人们的关注。

心理学家奥尔波特曾在《偏见的本质》中提到, 偏见是大脑偷懒的常态, 算法延续人类固有的偏见亦是如此。如今人类制造数据的速度惊人, 2018 年全世界创建等的数据总量为 33 泽字节, 预计到 2025 年将达 175 泽字节。我国虽无肤色种族问题, 但性别、年龄等方面歧视冲突不断, 还有大数据杀熟、信息茧房等影响

生活。人类即将进入新社会, 出现一门全新得无法想象的社会科学, 科技不再单纯是生产力工具, 而其本身就是一门科学, 慢慢会变成一切规则本身, 我们需要不懈努力以完好把控利用它。

2 当前研究综述

算法歧视这个词并不新鲜, 它的历史几乎和算法一样长。目前的研究主要围绕在成因、分类、解决措施及相关领域。关于成因, 丁艺、邓榕 (2021) 认为由现实社会歧视延伸、错误数据及算法不透明性所致, 并按原因进行了分类。曹博 (2020) 对比算法与传统决策歧视结果分类。许丽颖等 (2022) 发现算法歧视道德惩罚欲低, 潜在机制是人们认为算法更缺乏自由意志。徐昊 (2021) 从消费者保护视角探讨大数据杀熟法律规制。沿着学者们开辟的研究路径, 我们得以站在更高视角, 对算法歧视进行更为细致的拆解。

2.1 基于来源之辨: 先存歧视与技术歧视的分野

先存歧视与现实相连, 其一源于个别算法设计者的身份歧视, 其设计交友程序会按自身意愿屏蔽申请。二则来自社会文化, 难以回溯源头, 悄然融入系统, 不易被使用者察觉。

技术歧视, 主要原因是计算机工具的客观局限性, 表现在以下两点。第一, 技术初期, 受硬件、软件或操作者水平限制, 牺牲部分人需求。第二, 技术成熟后, 不可控又会成为新的问题。我们都承认算法的复杂性, 虽然是人创造出来的, 即便是专业技能精湛的程序员,

也无法保证面对由自己熟悉的规则排列组合而成的新的庞然大物时，总能一针见血地指出问题。尤其是人工智能加速发展的时代，专业领域的技术会越加“自主且复杂”。可以想象，当算法系统被赋予了类人的归纳和学习能力后，技术歧视会以更快的速度席卷各个领域。

2.2 聚焦作用对象：群体与个体面临的歧视困境

有损群体包容性的歧视，是指向种族或特定环境下对妇女群体、弱势群体的忽视。如算法正义联盟创始人乔伊提出的编码凝视概念，算法在关于雇佣、贷款乃至罪犯的服刑期限的决定方面有了更高的权重，这会使人很容易置身其中麻木不仁，也无法确认自己正在受到何种侵权行为。

有损群体公平性的歧视，体现在可能出现的预测及决策不公，最常见的便是招聘领域。2018年，路透社报道亚马逊设计的AI招聘算法对女性求职者不利，尽管简历中并未涉及明性别、种族等个人隐私，但如果简历中有“女子象棋俱乐部队长”或“女子学院”这种词，该算法就对直接对求职者作降级处理，不看个人能力，不论工作经验，女性竟成了降级的触发条件。亚马逊最终关闭了这个算法，并解散了相关团队，警醒了后人。

有损个体利益的歧视，是指算法还可能会被用于利益竞争，拥有技术储备及市场优势的经营者更容易通过算法技术去设置贸易壁垒，破坏正常的市场竞争机制。例如谷歌PageRank的算法在对网页链接重要性进行评价时，会把Google Shopping中自己的商品置于网页的显眼位置，而不用和其他网页商品进行竞争。

3 算法歧视的原因：现实、技术与数据短板

纵观算法歧视的历史，不难发现它的原因。

首先是现实社会歧视在虚拟世界的延伸。算法是人工产物，开发者难避自身价值取向，其嵌入影响算法运行。“毋庸置疑，算法歧视在很大程度上不过是人类社会早已扎根的种种偏见与刻板印象在虚拟世界的投射。”正如美国麻省理工媒体实验室教授拉胡尔·巴尔加瓦所言：“有偏见的不是算法，是我们人类。”它作为一种工具，是人类意识形态的具体体现，为人所设计也为人所所用，所以植入的价值选择必然会导致内隐偏见。警醒了这一点后，我们便要有意识地审视自身，避免将有色眼镜戴到机器上。

接着，算法本身的不透明性导致算法歧视。人工智

能算法系统的复杂程度堪比“黑箱”，只输出平面的数字，使外人难以回溯和检验依据。普通用户只能被迫接受，增加了纠偏难度。

此外，错误的数据也是推手。“垃圾进，垃圾出”，数据质量决定算法输出。若数据不佳，输出有偏差，算法还会沾染陋习。普林斯顿大学Aylin Caliskan等学者使用内隐测试量化人类偏见时发现，利用网络常见语言对系统进行训练时，机器学习程序自动从文本语料库中推导的语句包含着类似人类的偏见，可见被赋予了自下而上学习能力的算法并不能鉴别偏见和歧视。许多著名的虚拟社交机器人都曾有过这种轶事，如苹果系统的Siri在对话中学到新词汇和不甚规范的用语，百度搜索引擎小冰会模仿提问者的说话风格。或许有的个案看起来过于极端，但无法回避的是，机器学习的确需要大量数据才能完美理解，它的错误实践甚至在特定场景下改变部分人的命运。一名少数族裔表示，机场的CLEAR系统经常识别不到自己，所以不得不进入涉及背景和指纹调查的TSA Pre-check程序，不仅耽误时间还带来了时间、金钱和情感负担。

当有问题的算法被应用于司法程序，哪怕极轻微的数据偏差也有可能造成司法不公，这微小如尘埃的偏差落到一个具体的人身上，就是一座大山，积累的后果会严重破坏司法公正。

4 多维破局：从宏观到微观驯服算法偏见

面对算法失控，各界共识是让算法遵循伦理准则。

宏观层面，公平正义是全社会共同的价值追求，算法设计者、运营者和使用者应共同秉持原则，将算法公平内嵌到人机互动的整个系统之中，落实到加快技术迭代、完善制度和平等参与等。

中观层面，许多经验和案例研究已经证明，人类意志始终贯穿其中，并且还可能生成新的“机器偏见”。因此，应将公正性原则作为全过程的指导原则，并根据具体实践丰富它的内涵和维度。

微观层面，可从技术和制度两个维度分论。就技术维度而言，可通过引入纠偏工具、把控数据源头、优化数据质量等方法落实算法在技术维度上的治理。国际上许多企业已采取了有关措施，如2018年5月，Facebook公开Fairness Flow以检测和排除由肤色、性别和年龄等因素导致的不公平判断，用自动预警提醒开发者；2018年6月，Microsoft联合外部专家，筛查、校正和

拓展了用于训练 FaceAPI 的数据集,对肤色、性别和年龄等要素在数据集中占的比例进行调整,从而改进了分类器。就制度维度而言,监督行业协会的自律管理,开展从业人员伦理理念培训,建立科技伦理委员会等方式都有助于制度治理,加强对具体模块的治理及问责。由此,具体落实到操作层面,平台或企业需要做到以下几点:第一,保证数据公正,兼顾数据和决策主体的可见性和自主性,守护老幼人群、残障人士及其他弱势群体的利益。第二,坚持“算法透明性”,做到信息的公开、易获取和可理解性。第三,建构多元主体的算法问责体系,用群策群力提升技术风险治理能力。

5 总结

算法是现代科技的重要构成要素。随着算法歧视与价值伦理问题浮出水面,“科技向善”融入企业乃至社会发展的结构之中,是大势所趋人心所向。算法本身并无原罪,人文与技术的理性相容交织,才是可信赖可持续发展路径。我们应当认识到,技术发展方向和人类文明进程不应是冲突的,各界力量要联合起来,不断优化技术的同时,共建透明可视化的监督体系,从现在开始摒弃偏见,塑造健康包容的智慧社会,持续创造新的共享价值。

参考文献

[1] 曹博. 算法歧视的类型界分与规制范式重构[J]. 现

代法学, 2021, 43(04): 115-126.

[2] 丁艺, 邓榕. 算法歧视的伦理问题及其治理[J]. 科技传播, 2022, 14(20): 7-11.

[3] 刘培, 池忠军. 算法歧视的伦理反思[J]. 自然辩证法通讯, 2019, 41(10): 16-23.

[4] 许丽颖, 喻丰, 彭凯平. 算法歧视比人类歧视引起更少道德惩罚欲[J]. 心理学报, 2022, 54(09): 1076-1092.

[5] Caliskan, A., Bryson, J. J., & Barayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 183.

[6] CNNIC. (2024, September 19). 第54次中国互联网络发展状况统计报告. 199IT 互联网数据资讯网. Retrieved from <https://www.199it.com/archives/1718242.html>.

[7] Hindustan Times. (n.d.). Fairness Flow: Facebook Builds New System for AI Neutrality. Retrieved from <https://www.hindustantimes.com/tech/fairness-flow-facebook-builds-new-system-for-ai-neutrality/story-BDkWK0rDDnaywuDVJokB9N.htm>

作者简介: 叶慧莹(1998-), 女, 汉族, 福建宁德人, 硕士, 研究方向: 人力资源管理。