● 韩　勇

算法歧视的成因及治理路径

摘　要：随着人工智能应用的不断深入，算法歧视派生出了新的现象，其成因也更加多元。治理算法歧视问题，主要有伦理路径、技术路径、法律路径，要具体问题具体分析。

关键词：算法歧视；人工智能；大数据；数据瑕疵；技术缺陷
DOI:10.15997/j.cnki.qnjz.2021.08.048

“算法歧视”来源于大数据技术的应用，大数据技术会根据浏览记录、阅读喜好、购买记录等数据分析出用户的隐形特质。由于算法歧视根植于程序之中，其歧视行为也更加隐蔽，特别是在算法黑箱现象泛滥及数据正义缺失的情况下，人们很难发现歧视的存在。目前，算法歧视已被列为人工智能技术应用不可忽视的风险之一。

算法歧视的成因及具体表现

算法歧视的成因主要包括“数据缺陷”和“技术缺陷”两方面，但随着人工智能应用的不断深入，算法歧视也派生出了新的现象，其成因也更加多元。据此，算法歧视主要有以下四种表现方式。

1. 算法歧视的偏见表现

算法歧视的偏见表现是指程序设计者将自身偏见引入算法中从而引起的歧视现象。例如，在2018年玛德西亚杯游戏比赛中来自中国的战队IG最终夺冠，为了庆祝IG夺冠王思聪在微博上进行了抽奖活动，在最终113名获奖者中只有一位男性用户，而官方数据显示，在参与抽奖的用户中，男女比例为1:1.2，性别比并不存在悬殊差异。于是有许多网友质疑新浪微博的抽奖算法，甚至有用户主动对算法进行测试，将获奖人数设置为大于参与人数，但依然有大量用户无法获奖。据业内人士推测，在参与抽奖的用户中有一部分人很少发布原创微博，一般都是转发，此种行为很可能被算法判定为“僵尸号”或“机器人”，在未来任何抽奖活动中都不可能中奖。正是由于新浪微博抽奖算法的偏见性，一时间微博“算法事件”引爆网络。

又如，2018年2月加纳裔科学家Joy Buolamwini偶然间发现人脸识别软件竟无法识别她，除非她带上一张白色面具。出于好奇，她发起了“Gender Shades（性别阴影）”研究，发现人脸识别产品针对不同性别、不同肤色的人时会产生偏差，浅肤色男性错误率为0.8%，而深肤色女性的错误率则高达34.7%，研究者将这种偏差称为“算法偏见”。

此外，在搜索领域也存在着“算法偏见”现象，例如，在百度图片中以CEO、老板、总裁为关键词进行搜索，搜索结果大多是男性图像，女性图像寥寥无几。

可见，“算法偏见”的出现是因为算法设计者将自身偏见融入算法中，由此算法也具有了设计者的思维模式。所以，“算法偏见”可看作人类偏见在人工智能领域的折射，其本质仍然是人类固有的偏见思维和标签思维。

2. 算法歧视的数据瑕疵表现

算法歧视的“数据瑕疵”表现是由于所收集的数据存在瑕疵导致的。例如，世界上许多互联网企业都是男性工作者居多，技术性岗位更是如此，这就导致互联网企业在利用算法筛选数据时难免会出现性别歧视。这是由于“简历筛选算法”是根据企业现有员工构成进行训练学习的，其中包括性别、年龄等因素。由于“老师”（员工构成）所含有的男性员工较多，所以“学生”（筛选算法）在筛选时则会优选男性应聘者，从而导致性别歧视的出现。

另外，“数据瑕疵”还表现在数据来源的不统一上，人所共知的是，算法所学习的数据越多，产生错误的概率也就越小，而且其结果也将更加精准，但实际上并非如此。这是因为主流项目永远拥有更多的数据量，而非主流项目则少之又少，这就导致数据量多的一方结果更加精准，而数据量少的一方错误概率则会更高。例如，一名用户同时测试《红警》和《英雄联盟》两款游戏，而这位用户在此之前均未涉猎过与之相似的游戏，假设每天在相同的生理及心理状态下进行同等时间的训练，那么该用户这两款游戏的水平应当是稳步提升的。但经过一段时间之后我们会发现，用户在玩《红警》时玩法模式较为单一，而玩《英雄联盟》时则会使用许多不同的战术。这是因为，《红警》这款游戏早已过时，用户在训练时只能一遍一遍和电脑或极少的人类玩家进行对战，用户水平很快会达到瓶颈。而《英雄联盟》是时下最为火爆的游戏，每天都有无数的人类玩家与之对战，每天都能学到新的战术，在对战中能够十分清楚地感受到水

平的提升。所以，当两套算法相比较时，数据多的一方其结果更为精准，虽然可能在学习的初始阶段两套算法差距不大，但久而久之还是会呈现出较大差异，这就是数据来源的不匹配性。

3. 算法歧视的技术缺陷表现

算法歧视的技术缺陷表现是指，算法本身是中立的，但由于在技术设计方面存在缺陷从而导致歧视现象的产生。如果说“偏见表现”是主观意愿所导致，那么“技术缺陷”则是无意为之。以世界上最大的图像识别数据库 ImageNet 为例：该数据库中许多图片均被用户手打注释贴上了各种细分标签。尽管我们无从调查这些贴标签的用户是否带有各种偏见，但他们的确定义了“失败者”“妓女”“罪犯”“无害”的样貌。而这些标签在设计之初是不存在的，是被用户贴上的，这就是技术的缺陷所导致的。

此外，2016 年微软上线的 AI 聊天机器人 Tay 最初是以一个清纯可爱的少女面向世人，“她”最初的设定是根据周围环境来进行学习。也就是说通过不断与用户的对话来丰富自己的语料库。但很快被个别网友充满种族歧视、性别歧视的语言“带坏”，彻底变为一个带有各种社会偏见的集合体，上线仅仅一天就被迫下线。

4. 算法歧视的利益驱动表现

算法歧视的利益驱动表现是指，算法使用者为了追求自身利益最大化从而选择对自身最为有利的算法，而忽略了算法的公正性和合理性。利益驱动主要表现为“价格歧视”，即电商平台利用大数据收集用户支付数据，从而计算出用户愿意为该商品支付的最大金额，此种现象通常被人称为“大数据杀熟”。

算法歧视现象的治理路径

1. 算法歧视治理的伦理路径

目前持有这一论点的研究者各有各的侧重点，但他们普遍赞同将“算法透明”作为算法准则融入大数据的设计和具体应用中。这也意味着算法设计者首先需要公开自己的设计意图和运行机制，接受社会的监督。其次，还要将“反数据歧视”纳入算法中，使每个人都能在算法中获得均等的机会。再次，还要建立起评价机制对算法中出现的歧视现象进行具体评价，以便后续进行更正和管理。最后，算法设计人员还要具备“以人为本”的精神，铭记算法被制造被训练的目的不是为了榨取民众钱财而是应当给予民众生活便利，同时，算法设计人员还应提升自身职业素养与道德素养，在设计算法时尽可能避免将自身偏见代入算法中。

2. 算法歧视治理的技术路径

这种治理路径实际上是秉承“解铃还须系铃人”的原则，算法歧视的出现归根结底还是技术层面出了问题，既然是技术上的问题，就应当交给技术来解决。当前，许多研究者正在积极研发各种技术工具，旨在不降低算法精准性的同时，尽可能避免歧视现象的出现。例如：由数据科学家 Been Kim 所在团队研发的“概念激活向量测试”（Testing with Concept Activation Vectors）技术，简称为“TCAV”，该技术在 Google I/O 2019 大会上面世，它是一种算法可解释性的方法，能够直观地显示算法运算所依据的概念及其比重，能够观察其他 AI 模型中的算法倾向，而且该技术能够将机器学习模型变得更加通俗易懂，易于人类理解。比如，人工智能技术可以检测一个图像中的动物是否是老虎，而 TCAV 的作用则在于将人工智能识别过程细分化，帮助人类理解在识别过程中哪些变量发挥了作用，以及各自发挥了多大的作用，将人工智能的运作机理完全展现在人类眼前。因此，TCAV 这一功能特别适用于对算法的纠偏，当这项技术用于针对社会现实模型的识别时，就能够很容易地判定该算法是否有歧视。

3. 算法歧视治理的法律路径

对于算法歧视的法律治理，首先应该规范算法的使用范围、方式和底线。对于算法歧视，既要严厉惩处滥用算法的企业，又要考虑对由“算法歧视”所带来的侵权现象的赔偿规则。另外，还要界定算法的使用底线，对那些高度敏感的问题应当禁止算法介入，比如种族、民族、宗教等。此外，还要赋予用户更多的拒绝权利，比如：用户可以选择同意客户端收集其个人信息，如果用户认定算法存在歧视，那么用户有权拒绝其处理结果。

另外，对于算法的规制不能总是“马后炮”。应当运用法律的强制性，在算法设计之初就将人文伦理、算法透明、算法解释嵌入其中，如果算法出现歧视现象，还要启动问责机制。此举能够从法律层面约束算法设计者，使其谨守法律底线和伦理规制，避免“算法歧视”的出现。

而对于那些由利益驱动的算法歧视来说，“大数据杀熟”实际上是由算法主导的差别定价，虽然此举谈不上违法，但如果不对价格进行明示，则涉嫌侵犯消费者的知情权。同时，网信办还要出台相关措施来规范电商企业使用算法的行为，将算法圈在法律的框架内。另外，在当前的人工智能时代，消费者也要提升自身维权意识，在遭遇价格歧视时第一时间保留好证据，以便后续利用法律来维护自身合法权益。

参考文献：

[1] 张爱军，李圆．人工智能时代的算法权力：逻辑、风险及规制 [J]. 河海大学学报（哲学社会科学版）.2019(6).

（作者为武汉传媒学院新闻传播学院讲师）