抗击算法偏见

(上海交通大学　卢宏涛　编译自 Julianna Photopoulos. *Physics World*, 2021, (5): 42)

2011年，在佐治亚理工学院读本科期间，计算机科学家Joy Buolamwini发现，她与机器人一起玩躲猫猫这个简单游戏是不可能的，因为这个机器人不能识别她黑肤色的脸。2015年，已是麻省理工学院媒体实验室硕士生的她也在人脸分析软件中遇到了类似的问题：只有当她带上一个白色面具的情况下，系统才能检测到她的脸。这是巧合吗?

Buolamwini的好奇心驱使她在4个人脸识别软件中尝试识别自己的图像，她发现，软件要么完全不能认出人脸，要么把她的性别识别错。然后她决定用来自3个亚洲和3个欧洲国家的政治家的1270幅人脸图像进行测试，这些人脸图像具有不同的特征、肤色和年龄。结果发现这些人脸识别技术有几乎35%的几率会错误识别深肤色女性人脸，而对白人男性的识别正确率堪称完美(99%)。

纽约大学AI Now研究所的Joy Lisi Rankin说："计算机是由人编程的，而人即使心怀善意，仍然可能会被带偏并怀有歧视"。

物理学家正不断地在各种各样的领域中运用人工智能(AI)和机器学习(ML)。来自费米国家加速器实验室的物理学家和数据分析师Jessica Esquivel说："作为粒子物理学家，我们的主要目的是开发算法和工具，帮助寻找超越标准模型的物理规律。我们没有预见这些算法和工具可能被布署到技术中，应用于日常社会中进一步压迫边缘人群"。他目前的工作是开发AI算法以增强μ介子g-2实验中的粒子束存储和优化。

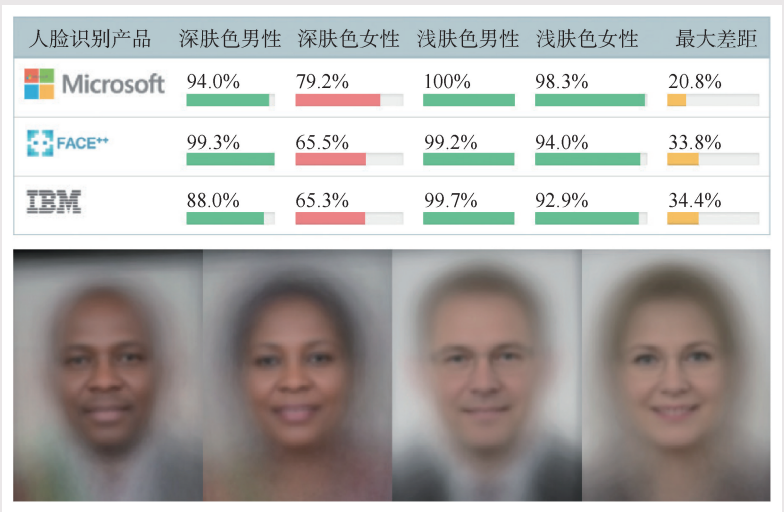
普林斯顿大学粒子物理和机器学习研究人员Savannah Thais说："数据缺乏多样性也影响了开展的研究工作和开发的系统。"一个例子是亚马逊实验性的招聘算法，该算法基于他们过去的招聘实践和申请人数据。亚马逊最终放弃了那个工具，因为性别偏见从过去的招聘经验中太深地嵌入到他们的系统中，倾向于拒绝女性求职者，不能保证公正性。

在物理学中具有多样性是极其重要的，Thais目前在欧洲粒子物理研究中心(CERN)为高光度大型强子对撞机开发加速机器学习重构算法。他说"大部分物理研究者没有与其他种族、性别和社团人群直接生活的经验，而他们正是被这些算法影响的人群。"

美国斯坦福大学人工智能研究人员Pratyusha Kalluri去年在*Nature*上发表文章写道："是时候将被边缘化和被影响的人群置于AI研究的中心了——他们的需求、知识和梦想应该引导着算法研发。"

物理学家的作用

费米实验室的Brian Nord是一名宇宙学家，利用AI搜索宇宙起源和演化的线索，他解释到："望远镜在多年的巡天中扫描天空，采集了包括图像的大量复杂数据，我们用AI技术分析数据，寻求理解暗能量，其引起了宇宙的加速膨胀。"但是在2016年，他认识到AI可能会有偏见。有一款称为COMPAS的风险评估软件，它被应用于美国法庭上预测哪些罪犯最有可能再犯罪，由此决定设置保证金数目。调查发现，不管犯罪的严重性及真正重犯的可能性，黑人比白人以几乎两倍高的概率被标记为高风险。



三个公司的人脸识别产品对1270幅图像识别结果的比较

Nord组建了一个物理学家和计算机科学家联盟，致力于在开发算法时争取更多的审查。他警告说："物理学家应该了解诸如数据隐私问题，数据和科学如何被应用于侵害公民权，技术如何被用于维护特权，数据驱动的技术剥夺了人的权利等问题。"

为了使这个问题引起更广泛的注意，Nord、Esquivel和其他同事写了一封信给整个粒子物理社团，信中讨论了"计算研究的伦理内涵和科学家的作用"，强调为何物理学家应该注意他们正在构建和实现的算法。Thais也敦促物理学家们主动地介入到AI伦理中，虽然一般情况下机器学习的物理研究应用"不会涉及到伦理问题"，但许多物理学家之后会进入计算机软件、硬件和数据科学公司工作。"很多这样的公司在使用人类数据，因此我们必须让我们的学生以负责任的方式做那些工作"，她说。

Thais和Esquivel都相信物理学家在理解和监管AI方面可以发挥重要作用。"面对一个更像黑盒的机器学习算法，我们真的想了解算法的精确度怎么样，它如何处理边缘情况，为何在特定问题上表现最佳。Thais说，"这些是物理学家以前做过的工作"。

算法决策工具也许是为科学研究开发的，但后来被用于商业监控场景，其中任何的数据偏差都会有现实后果

审计算法

2020年英国数据伦理和创新中心发表了一篇关于算法偏见方面的评论，发现在过去的几年中，招聘、金融服务、警务和地方政府等部门应用算法进行决策有明显的增长，并发现了算法决策存在偏见。这个报告要求各类组织、机构主动地利用数据来辨识和减轻算法偏见，并了解算法的能力和局限。

数学家Cathy O′Neil在2018年成立了一家咨询公司，与其他公司合作并审计他们的算法。Buolamwini也试图通过她的非赢利算法正义联盟(她于2016年成立的跨学科研究机构)创建更公正和负责任的技术，以了解AI技术的社会影响。2018年她与计算机科学家Timnit Gebru一起对前述涉及算法偏见公司的后续研究进行重新审计，并增加了亚马逊和Kairos两个公司。研究发现亚马逊的人脸识别软件居然不能准确地识别米歇尔·奥巴马的脸，但之前3个公司的系统已经有很大改观，说明他们的训练数据集已经包含了更多样的图像。

这两个研究具有深远的现实影响，导致了两项美国联邦法案的颁布——算法问责法案和无生物识别障碍法案，以及纽约州和马萨诸塞州的州法案。这些研究也帮助说服微软、IBM和亚马逊暂停了将他们的人脸识别技术应用于警务。

2020年计算机科学家Deborah Raji与谷歌的同事一起开发了一个为AI追责进行算法审计的框架。"内部审计是必不可少的，因为它可在一个系统部署到世界各地前对其进行修改，审计在开发流程中可能引入偏见的那些环节是很重要的。"

消除算法偏见

2019年AI Now研究所建议研究AI偏见应该超越技术方向。"我们不仅仅需要改变算法或系统，更需要改变研究机构和社会观念"，Rankin解释说。她认为，为了清除或监管算法偏见和歧视，需要"大规模全体行动"。在这个过程中让自然科学界之外的人们参与进来会有帮助。

Nord同意物理学家应该与其他学科的科学家以及社会学家、伦理学者一起工作。"不幸的是，我没看到物理学家或计算机科学家与花费了大量时间和精力研究这些问题的其他领域研究人员充分合作"，他说到，"看起来每几周都会有一个新的有害的机器学习应用试图做有偏见的事情"。例如，德克萨斯大学奥斯汀分校直到最近才停用了一个用来预测研究生是否能成功完成学业的机器学习系统，它的数据是基于之前的入学数据，可能有偏差。"为什么我们要在必然的一个人本主义空间中追求这样的技术官僚解决方案?"Nord发问。

Thais坚持认为物理学家必须对这些偏见问题的现状有更好的认识，然后了解别人为减轻这些问题采取的努力。"我们必须把这些对话带入到关于机器学习和人工智能的所有讨论中"，她说。

Nord甚至认为，"在问物理学家他们是否可以之前，应该问他们是否应该创建或实现某种AI技术"，他同时补充，停用现存的有伤害的技术也是可行的。"这些技术的使用是我们作为个人和作为社会做出的一种选择。"