DOI:10.13811/j.cnki.eer.2021.08.002

人工智能教育应用的偏见风险分析与治理

沈 苑，汪 琼

（北京大学 教育学院，北京 100871）

[摘 要] 在倡导教育包容与公平的时代背景下，教育面临着以个性化、自适应为特征的智能技术所带来的偏见风险。在智能教育产品的研发过程中，潜在的设计偏见、数据偏见和算法偏见会使教育领域面临多重风险，如教育活动的公平性受阻、教育系统内部的歧视加剧、教育中的多元化特质消解和学生思想窄化。学校、科技公司、监管机构等组织应当携手共进，提前侦测偏见风险并加强治理，包括：提升研发团队的教育理解和多元眼光，让师生成为产品的合作创造者，加强偏见检测和道德评估过程，建立可解释、可审查、可纠正的系统优化机制，开展面向师生的人工智能素养教育，确立人工智能在教育中的应用范围和伦理规范，从而争取实现人机良性互动，打造多元包容的智能教育环境。

[关键词] 人工智能教育应用；人工智能伦理；教育包容；算法偏见；数据偏见

[中图分类号] G434 [文献标志码] A

[作者简介] 沈苑（1996—），女，江苏苏州人。博士研究生，主要从事人工智能教育应用伦理研究。E-mail：shenyuan@pku.edu.cn。

一、引 言

当前，关于人工智能教育应用（Artificial Intelligence in Education，简称 AIED）的讨论非常热烈。无论是作为辅助教学、校园管理的工具，还是学科教育的重要内容，人工智能都为教育发展注入了新动力。有支持者主张，AI 是基于真实教育数据做出决策的，较之人类决策更加客观、公正和权威。然而，真的是这样吗？普林斯顿大学的一项研究发现，AI 和人类一样带有偏见，它们会将女性群体与“家庭”“艺术”挂钩，将男性群体与“事业”“野心”挂钩，将有色人种群体与一些令人不悦的词语挂钩[1]。随着 AI 在各行业中承担起愈发重要的决策角色，诸如此类的偏见足以对我们的真实世界产生影响，教育领域也不例外。

社会心理学家将“偏见”定义为对某一社会群体及其成员的一种不公正态度，是一种事先或预先就有的判断[2]。如果教育决策含有偏见，可能会阻碍学生的自由发展，加剧学生之间的差距，不利于教育公平和包容的发展。20 世纪的美国曾经一度施行“红线政策（Redlining）”——依据居民的职业、种族、地区、房屋状况等指标划分居民区等级，被划入红线区就意味着无法获得住房贷款，有色人种和贫困人群绝大多数都在红线区内[3]。此政策原本是为了高效分配社会资源，却导致了更加严重的歧视。在教育中，学生性别、健康状况、年龄层次、经济条件、家庭住址等都是隐形的“红线”，可能导致优势学生和弱势学生这两个群体的固化甚至相互隔离。

本研究基于真实案例对 AIED 偏见研究的背景展开讨论，结合智能技术和教育场景的特点分析偏见的生成机制和面临的风险挑战。本研究结合当前已有实践所提出的风险治理的具体策略，对于推动 AI 赋能教育、提升教育包容、促进教育公平具有重要意义。

二、研究背景

有学者指出：“‘适应与个性’和‘包容与公平’都是现代教育技术发展的重要范畴，但令人惊讶的是，却鲜有这两个领域交汇处的研究。”[4]AIED 及相应的创新教育模式一方面能够促进优质资源共享，弥合教

基金项目：教育部—中国移动科研基金 2018 年度项目“中小学编程教育与人工智能工程素养研究”（项目编号：MCM20180611）

育鸿沟,推动教育公平进程;但另一方面,因社会权力结构分布带来的偏见,可能会因为对适应性教学和个性化学习的追求而被巩固和加剧。如今,我们站在这个十字路口,应当厘清教育包容与公平和适应与个性之间的关系,充分认识研究 AIED 中所存在的偏见问题的意义。

(一)智能时代下教育包容与公平的新局面

包容与公平始终是教育的核心主题。联合国教科文组织(UNESCO)将“确保包容和公平的优质教育,让全民终身享有学习机会”作为 2030 年教育分支的可持续发展目标之一[5]。经济合作与发展组织(OECD)将教育公平和教育包容作为衡量教育成功的两项标准[6]。当 AI 进入教育场景,教育包容与公平的内涵也必将迎来新的变化。只有在确保 AIED 不会阻碍教育包容与公平的前提条件下,才能真正可持续性地为教育赋能。

包容与公平不仅意味着使用 AI 时要涵盖有特殊需求的学生,还意味着为所有学生群体提供平等的、可获取的、高质量的智能教育支持。AI 使用是否会导致受教育机会不均等,AI 能否为教育环境中的所有主体提供服务,AI 是否根据用户的健康状况、性别、年龄、家庭社会经济地位、文化背景等因素提供有差异的服务,AI 是否会给某些学生群体贴上污名化标签,学生是否会因为 AI 的介入而感到不被重视和不被尊重?种种问题亟须明晰的答案。

(二)个性化支持与教育歧视的一线之隔

实现个性化学习基本上是教育领域终极追求的目标[7]。过去数十年,教育技术越来越强调以学生为导向,推动个性化自适应学习的发展[8],如智能导师系统、教学机器人、智能学习伙伴、自动化评估系统、虚拟现实技术教育应用、个性化慕课等[9]。但是,当偏见悄悄潜入其中,“因材施教”可能因此变味,与“教育歧视”仅是一线之隔。

美国公立学校选择创新机构(Institute for Innovation in Public School Choice,IIPSC)开发出一套高中入学推荐系统,通过收集分析三类数据:父母希望孩子去什么学校、每所学校和每个年级的名额以及学校的录取要求和规定,从而为纽约、波士顿、丹佛等地区的学生推荐高中[10]。研究发现,此系统会迫使家长和学生在信息不透明、不全面的情况下做选择,而且弱势学生群体处境并没有因此而改善,他们大多被推荐去往外地或较差的高中[11]。而且,由于算法的“黑箱”特质,缺乏专业技术知识的学生、教师、家长和管理者难以对系统输出的结果进行审查。

三、AIED 的偏见生成机制

有学者指出,社会不公是由社会结构性壁垒和个人认知性壁垒构成的[12]。前者是指历史上的不公在政策、实践、价值观中的渗透;后者是指用于维持社会不公结构的个人层面的思考活动会潜在地加深理解、行为、决策中的刻板印象。这两层壁垒作用于 AIED,使之沿袭甚至强化带有偏见的社会思考模式,导致学习者被无限循环的偏见所包围。偏见会贯穿于 AIED 的生命周期——从系统的初步设计、收集数据、建立数据库、算法编写、模型训练,直至应用落地,可归纳为以下三种生成机制:

(一)设计偏见:特定群体被排除在目标受众之外

有报道称,尼康某型号的相机在拍摄亚洲人时总是会弹出“是否有人眨眼?”的提示。加纳裔美国研究员乔伊·布拉姆维尼(Joy Buolawini)在使用实验室设备的时候,屡次遭到人脸识别系统的“无视”,某天她带上了一张白色面具,系统居然成功识别了她的“面部信息”[13]。有些产品在设计之初就忽略了某些特殊群体的存在,致使某些特定学习者群体无法与其他学生同等获取系统的服务。有研究者指出,这是一个教育发展至今都未曾面临过的、全新的歧视维度[4]。

从理论上来说,随着 AI 的发展,学生在认知能力、文化背景、身心健康等方面的不同需求能够获得更充分的个性化支持。但目前绝大多数 AIED 系统面向的仍是一般学生群体,鲜有涵盖特殊学生群体的特定需求。某些基于眼动追踪技术的智能系统,无法准确分析视力障碍、阅读障碍、注意力集中障碍学生的眼动模式。虽然使用开放学习者模型来提升个性化教学质量的做法日渐普遍,但学习者要有一定的认知和技术能力,才能有效参与这类复杂模型的构建与控制[14],患有认知障碍的学生将会面临巨大的困难。如果产品设计初期就未将特定群体的需求纳入考量范围内,在后续数据收集和模型建构过程中,他们很可能会被边缘化甚至被彻底无视。

(二)数据偏见:社会结构性壁垒的复制与重现

数据正在重塑着我们的教育,从课堂分析到成绩预测,从招生入学到生涯规划等,越来越多的教育活动在不断被数据化。数据科学家维克托·迈尔-舍恩伯格(Viktor Mayer-Schönberger)指出,大数据和教育的结合将超越过去那些“力量甚微的创新”而创造真正的变革[15]。用于训练、学习、挖掘的原始数据是 AI 系统的基石,其中立程度直接影响系统输出的结果[16]。

技术哲学家指出,人们对于技术经常会产生正向

偏见(Positivity Bias)[17],即在没有足够信息的情况下,默认信赖技术,对其客观程度过分乐观。“人类历史被隐藏在数据集里,如果我们要用数据集来训练系统做出自动决策,需要深刻意识到这种历史的存在。”[18]因此,如果未经过严格审查和测试,AI 不能充当客观公正的决策者角色,而是会重现隐藏在数据集中的人类偏见[19]。如果数据样本不全、边缘化某些群体或隐含某种偏见,将会导致数据库被污染,无限循环与强化社会的结构性偏见[15]。

虽然绝大多数工程师都希望在系统设计时尽量避免嵌入偏见,但很多看似中立的数据其实已经受到了影响。如监督学习算法训练出的模型会反映出打标签者的偏见[20]。如某些贫困地区学生的邮编、电话、学校都会潜在反映出“贫困”特征。数据的类型也会影响客观性,如美国某些州用智能系统筛选需要社会救助的儿童,用到大量田野调查和自我陈述报告的数据,其中包含大量社会情感因素,依据目前技术发展水平,难以排除所有主观内容[12]。AI 与用户互动时还会习得人类偏见,如模拟 19 岁少女的微软聊天机器人 Tay,不到一天时间 Tay 就被“教坏”了,不但辱骂用户,还发表带有种族主义、煽动性和攻击性的言论[21]。

(三)算法偏见:认知性壁垒渗透于算法模型中

算法偏见意味着工程师在编写算法过程中将自身偏见渗透其中,导致程序会输出带有偏见的决策[22]。当我们希望通过算法模型来进行客观决策的时候,其实我们的价值观、信仰和偏见都已经渗透其中,算法就是“镶嵌于数学中的人类观念”[23]。

受新冠肺炎疫情影响,2020 年的英国高考无法如期举行,英国资历及考试评审局(The Office of Qualifications and Examinations Regulation,OFQUAL)采用了智能评分系统为考生打分。此算法模型将考生以往的个人成绩、排名以及教师对学生今年成绩的预测、考生所在考区三年来的综合成绩水平作为重点指标。如果某个考区三年内总体表现不佳,系统会“无视”此考区学生当年的努力,默认他们当年也会表现不佳。即使这个较差考区中有一位非常优秀的考生,他也很可能会受到算法偏见的影响,得不到自己应有的高分。虽然研发此系统的初衷是希望避免教师打分的主观性,但此模型将“考生所在考区的水平”作为重要变量,导致最终输出的结果带有严重偏见——经统计发现,全国 39.1%的考生的系统得分低于教师估分,尤其是位于较差考区的弱势学生群体被系统“打压”得最厉害。在私立学校就读的学生成绩却没有受到太大的影响,拿到 A 等级及以上的学生甚至比之前多了 4.7%[24]。

四、AIED 的偏见风险分析

随着 AI 逐步深入教育场景,背后的偏见问题在应用过程中渐渐暴露和发酵,社会争议性事件屡有发生。如果缺乏对偏见问题的敏感度,事前疏于评估偏见风险,可能导致出现范围更广、程度更深的负面影响事件。

(一)教育活动的公平性受到阻碍

随着 AI 在教育中的全方位应用,偏见风险会渗透在各类教育活动中。例如:在预测学生下一次统考成绩时,某些学生可能因为所在学区的成绩差而被系统认定将会表现不佳;一位学生正在回答简答题时,智能系统的自动纠正功能改变了她的遣词造句,甚至违背了她原本想表达的意思;AI 在与某学生交流时习得了性别歧视并在和其他学生的交流时复制重现……虽然上述情况也许尚未出现,但在未来一旦发生,就很可能引发社会关注,有损教育公平。

教师智能评估模型曾在美国引起风波。教师谢里·莱德曼(Sheri Lederman)教学能力出色,但纽约州所用的教师增量评价模型(Value-added Model,VAM)却为她 2013—2014 年的工作打上“无效”的标签,使她面临着被开除的风险,她也因此提起诉讼[25]。此模型会比较学生在中等水平教师指导下可能会取得的成绩和学生在某个教师指导下实际取得的成绩。前者是系统给出的预测成绩,后者是学生的实际成绩[26]。通过比较,系统会自动为教师打上“高效”“有效”“发展中”“无效”的标签。这些标签会影响教师的声誉、薪资和去留。关于这类智能评估系统的一大争议就是:教师评级是否应该取决于学生考试成绩的进步?如果某学生上一次考试已经取得高分,他在下一次考试中就几乎没有进步的余地。虽然莱德曼的学生在 2014 年的进步并不显著,但他们在 2013 年成绩优异,因此,不能武断地判定她的教学是“无效”的。技术开发者带有“教师付出和学生进步具有因果关系”的认知偏见,对教师工作的评估过度倚重于学生进步情况,致使一位优秀的教师险些被学校开除。

(二)教育系统内部的歧视加剧

以学生为中心的智能学习环境中,偏见可能会导致特定学生群体遭到排斥,在性别、文化、地域、知识水平、家庭经济水平等方面加剧群体之间的歧视和隔离。如果说,不透明的全方位数据采集是对学生隐私权的“一次伤害”,那在此基础上形成带有偏见的决策和歧视性干预,对学生所造成“二次伤害”的影响则更深更广。在“算法中立”的幌子下,虽然我们并非有意

制造歧视,但是最脆弱的社会阶层很可能会受到不公平的对待[27]。有学者发现了针对特定学生群体的价格歧视——在报名某权威的 SAT 在线辅导服务时,输入亚洲人社区的邮编,将会被收取更高的费用,高达其他地区学生的两倍[28]。

目前教育系统中存在的各种刻板印象,比如“女生不擅长 STEM 学习”“来自贫困家庭的学生成绩差”,可能会经过算法模型而进一步加剧,导致现实中歧视现象愈发严重。而且,目前大多数系统并没有考虑到残障学生的特定需求,其语言、表情、行为可能无法被正确识别。这些学生原本就缺少强有力的支持和发展机会,处在成长时期的他们还会遭遇被智能系统边缘化的孤独无助。并且,算法具有复杂性、隐蔽性和不透明性,他们甚至可能全然不知自己遭受歧视。如果这些学生持续受到系统“惩罚”,这个过程中会产生更多的带有偏见的新数据,进而加剧教育中的“马太效应”。

(三)教育中的多元化特质消解

教育是一个充满未知的过程,不同学生做出某种行为可能都有不同原因。如果技术开发者难以洞察学生行为背后的真正含义,仅凭自己对于教育的理解,带有偏见地将学生行为解释为现有模式的某种子集或变体,不仅会得出错误的分析结果,更会阻碍教育环境下的多元发展。每个学习者的知识背景、思考方式、学习风格、兴趣爱好和行为习惯都有自己的特点,AIED 的目的是达到“因材施教”,为不同的学习者提供适应化和个性化的学习内容和策略。然而,当算法模型误判了学习者的行为,无法正确理解和包容每个学生的特点,无法准确提供学生所需要的针对性支持,很可能造成教师理解狭隘化、学习内容同质化、教学策略单一化等问题。上述这一切都有悖于 AIED 的初衷。

(四)“过滤气泡”窄化学生思想

算法模型具有渗透性,“就像流行病学模型可以用来预测票房,垃圾邮件过滤器被用作检测 AIDS 病毒的工具”,社交媒体中常用的精准推荐算法也会被应用在个性化学习领域[22]。这种应用尽管可以为学生提供针对性支持,但很可能会形成“过滤气泡(Filter Bubble)”——以大数据和算法推荐为底层架构,根据学生的搜索结果或使用习惯进行纪录与分析,过滤掉与学生观点相左的或学生不感兴趣的内容,只给学生提供他们想看的内容,导致学生接收到的资讯被局限于某个范围内,造成认知上和意识形态上的分化与隔离。

“过滤气泡”概念的提出者伊莱·帕雷泽(Eli Pariser)察觉到自己在脸书(Facebook)上见到的与他政治立场不同的言论越来越少的情况。脸书所采用的推荐机制是会给用户推荐其好友所点赞过、分享过的类似内容,这致使用户被包裹在狭小的信息空间里,听不见异质的声音[29]。而且精准推荐机制恰好符合了人本身所具有的验证性偏见(Confirmation Bias),即在面临众多选择的时候,人们大多倾向于听取与自身观点一致的观点。当 AI 在教育中成了“回音壁(Echo Chamber)”般的存在,学生的验证性偏见将会被不断地印证和加剧,导致视野和思想不断窄化,价值观和言行举止可能会变得偏激甚至极端[30]。

五、AIED 的风险防治策略

智能教育产品的每个发展环节都需要关注偏见问题,组建团队、确定用户需求、产品设计开发、宣传推广、应用落地都须严格把关。跨文化方法应该被纳入政策制定、产品研发、团队建设等环节中。学校管理者和决策者、产品设计者和开发者、学生、教师、家长、科研工作者应当携手推进教育环境包容性和多元化的建设。

(一)提升研发团队的教育理解与多元眼光

如果技术人员对于教育过程不够熟悉,对于教育本质理解不足,缺乏情境体验和背景知识,就很可能基于错误或片面的教育认识在系统开发过程中嵌入自己的偏见,导致算法模型与教学实践脱节。首先,应该提升研发团队的教育理解力。智能教育企业应当注重培养员工对教育的正确认识,比如可以邀请教育学领域专家开展面向企业管理者和工程师的讲座、工作坊等培训活动,修正他们在教育方面的认知偏见。其次,研发团队同质化也是导致偏见的重要原因。因为人们往往难以察觉自身的内隐偏见,对于他人利益敏感度不足,同质化的研发团队更容易忽视与自身不同的群体需求,收集的数据也更多是来自同群体的样本。如开发图像识别系统时,非洲程序员的训练集里黑人照片更多,而亚洲程序员更多用黄种人的照片。因此,不少伦理专家都提出,组建研发团队时要关注成员在社会属性上的异质程度[4]。最后,还应该关注成员之间的学科异质性,构建问题解决的多元眼光。如地域偏见可能不会因为团队里增加了几位偏远地区的成员就被消除,但当团队中有计算机专家、数据科学家、教育专家、伦理学家等一系列技术和非技术角色时,就能够更全面地理解教育中的地域偏见问题,从而为寻找创新解决方案提供多元视角。

(二)让师生成为智能教育产品的合作创造者

互联网时代下“共同创造”成为汇聚各方力量的创

新模式,师生无疑是教育技术重要的合作创造者[31]。智能教育产品的最终目标是帮助解决教育领域的问题。因此,企业应鼓励师生群体充分参与产品研发,积极与企业人员对话,清楚地传达需求和困惑。企业应对典型群体或特殊群体的需求展开充分的调研,邀请具有典型性或者特殊性的教师代表和学生代表,通过问卷、访谈、课堂观察等方式洞察用户行为,建立起鲜明的、动态的师生用户画像,在此基础上推进研发进程。

在 AI 全生命周期中,师生与企业应该始终保持互助互惠的关系,师生代表可以参与产品测试和试点使用,提出反馈建议帮助企业改进优化产品,使研发出的智能教育产品能直击教育中的痛点问题,为更广大的教师和学生群体提供有效支持。

(三)加强偏见检测和道德评估过程

改善 AIED 的偏见问题,需加强对数据集、算法和模型的偏见检测。检查用于训练系统的数据集是否包含了足够多的样本量,是否同质化过于严重,是否存在类别不平衡的现象,是否缺少某特定群体的数据样本,是否呈现出某种倾向性……如果侦测到了数据集存在着偏见倾向,要及时加以干预,比如重新选取分布更合理的数据源、修正数据比例、调整数据精度。同时,也要检查算法是否在设计之初就包含着某种特定偏见,是否会系统性地忽视或低估某些学生,模型是否通过了某些没被包含在训练样本里的特定学生群体的测试。谷歌开发的 What-If、IBM 开发的 AI Fairness 360、芝加哥大学数据科学和公共政策中心开发的 Aequitas、微软开发的 InterpretML 等偏见检测工具都可以被整合至模型开发过程中,帮助规避系统正式运行过程中的偏见与歧视。

同时,也需要加强道德评估过程。由数据科学家、教育家、技术专家、教师和学生等共同组成道德评估小组,对技术开发、试用、研究过程中可能存在的道德问题开展评估工作,帮助发现隐藏的错误认识或者偏见倾向。道德评估小组可以标记出数据抓取、转化、训练、预测、呈现决策等每个环节可能存在的偏见隐患。

(四)建立可解释、可审查、可纠正的优化机制

由于系统训练和验证数据集的过程往往是不可见的,最终产生的数据模型常常缺乏可解释性。如果要让带有风险的技术工具参与教育决策,应当建立起可解释、可审查、可纠正的系统优化机制,允许用户预览和更正决策。让教师和学生参与算法的形成、修改过程以及决策过程,是提高算法透明性和可解释性的重要途径[28]。

假设某学校采用智能分班系统对全年级学生进行分班,每个学生应该对自己的数据隐私和身份标签有较清晰的了解,他们有权利知道数据收集的范围和进行学生画像的要素,知道为什么自己会被分到这个班,并且他们应该有途径可以反对或纠正这项决策[32]。参考垃圾邮件筛选器,用户可以自主检查邮件过滤是否准确,也可以将被系统误判的邮件重新归类。由于师生不是专业的技术人员,应该为他们提供最为方便快捷的反馈路径,比如在 AIED 产品外形上或屏幕界面上设置清晰可见的撤回、预览、标记错误等功能按钮。

(五)开展面向师生的人工智能素养教育

当 AI 进入教育场景,如果教师不了解其特征和工作原理,可能会对技术产生过度依赖,盲目相信运算结果。当不透明的智能算法取代了教师对教学过程的自主判断,教师也许会面临着教学技能的退化和因材施教能力的丧失。如果学生不了解 AI,不熟悉和 AI 的相处模式,很可能会对他们的学习产生负面影响,阻碍学生的技术观和亲社会行为的发展,某些学生甚至会遭受 AI 的歧视。通过开展 AI 素养教育,帮助师生意识到 AI 并不是绝对正确的,它也像人一样存在着种种偏见,需要始终保持警惕,始终保留自己独立思考和自主判断的能力。引导师生不盲目迷信和依赖 AI,理解 AI 是为教育赋能的一种可选择的方法,并不是唯一路径。

(六)确立人工智能在教育中的应用范围与伦理规范

AI Now 组织曾建议:"'高风险'的核心公共机构,如司法机构、医疗机构、教育机构不应该使用不透明的 AI 算法。"[33]相较于彻底禁止教育中应用 AI,我们更应该限制 AI 在教育中的应用范围。我们不应在教育中模糊 AI 和人类的边界,不能任由 AI 一步步扩大它的控制范围。AI 渗透教育场景的深度、运算结果对决策制定的影响程度、在支持学生学习中扮演的角色,都需要明确规定和审慎决策。虽然对于应用范围的限制并不能完全消除偏见,但至少能够减少偏见对教育活动的渗透,避免社会偏见在技术中无限循环和放大。

严格的监督机制和明确的伦理规范也应当建立起来。学校、企业、科研机构、监管部门以及第三方组织应当商讨各方都接受且受益的监管途径,在已有的 AI 伦理准则基础上,建立起可信赖的 AIED 的具体原则规范,树立严格的规章制度以惩处应用中的偏见和歧视现象。

六、结　语

迄今为止,国内 AI 在教学中的应用还没有达到"广泛"的水平,绝大多数智能教育产品都还处于弱

AI 的范畴,其中的偏见倾向尚不足以造成严重后果。但也许在不远的未来,AI 将会获得更大的能力,足以左右孩子的教育和成长,控制人类的生活与思想。在感叹着 AI 为教育带来深刻变革与全新机遇,享受着智能化的服务与支持的同时,我们应当时刻保持着对于潜在“红线”的敏感和警惕。正如哲学家罗伯特·所罗门(Robert Solomon)所说:“我们曾经建立起来的那些关系,最终将会被用来塑造我们。”[34]

[参考文献]

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

[2] 张中学,宋娟.偏见研究的进展[J].心理与行为研究,2007(2):150-155.

[3] 王旭.“红线政策”与美国住房市场的反歧视立法[J].社会科学战线,2016(5):89-98.

[4] KNOX J, WANG Y, GALLAGHER M. Artificial intelligence and inclusive education[M]. Singapore: Springer, 2019.

[5] World Educators Forum. Education 2030: incheon declaration and framework for action towards inclusive and equitable quality education and lifelong learning for all[M]. Paris: United Nations Educational, Scientific and Cultural Organization, 2015.

[6] 刘宝存,屈廖健.PISA 2012 教育成功国家和地区的基本经验[J].比较教育研究,2015,37(6):14-20,29.

[7] 肖睿,肖海明,尚俊杰.人工智能与教育变革:前景、困难和策略[J].中国电化教育,2020(4):75-86.

[8] 祝智庭,魏非.教育信息化 2.0:智能教育启程,智慧教育领航[J].电化教育研究,2018,39(9):5-16.

[9] YU H, MIAO C, LEUNG C, et al. Towards AI-powered personalization in MOOC learning [J]. (npj) Science of learning, 2017,2(1): 1-5.

[10] HEROLD B. Custom software helps cities manage school choice [EB/OL].(2013-11-03)[2021-01-26]. https://www.edweek.org/ew/articles/2013/12/04/13algorithm_ep.h33.html.

[11] NATHANSON L, COTCORAN S, BAKER-SMITH C. High school choice in New York city: a report on the school choices and placements of low-achieving students[R]. New York: Research Alliance for New York City Schools, 2013.

[12] CAPATOSTO K. Foretelling the future: a critical perspective on the use of predictive analytics in child welfare [R]. Columbus: Ohio State University, 2017.

[13] ROSE A. Are face-detection cameras racist?[EB/OL].(2010-01-22)[2021-01-26]. http://content.time.com/time/business/article/0,8599,1954643,00.html.

[14] 危怡,胡梦华,胡艺龄,顾小清.开放学习者模型:让学习者参与构建——访国际知名教育人工智能专家朱迪·凯教授[J].开放教育研究,2018,24(3):4-11.

[15] 维克托·迈尔-舍恩伯格.与大数据同行——学习和教育的未来[M].赵中建,张燕南,译.上海:华东师范大学出版社,2015.

[16] 郭小平,秦艺轩.解构智能传播的数据神话:算法偏见的成因与风险治理路径[J].现代传播(中国传媒大学学报),2019,41(9):19-24.

[17] LIN P, ABNEY K, JENKINS R. Robot ethics 2.0: from autonomous cars to artificial intelligence [M]. New York: Oxford University Press, 2017.

[18] ROSENBERG S. Why AI is still waiting for its ethics transplant[EB/OL]. (2017-01-11)[2021-01-26]. https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/.

[19] MARCINKOWSKI F, KIESLICH K, STARKE C, et al. Implications of AI (un-) fairness in higher education admissions: the effects of perceived AI (un-) fairness on exit, voice and organizational reputation [C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York: ACM, 2020: 122-130.

[20] MITTELSTADT B D, ALLO P, TADDEO M, et al. The ethics of algorithms: mapping the debate [J]. Big data & society, 2016, 3(2): 2053951716679679.

[21] LEE D. Tay: microsoft issues apology over racist chatbot fiasco [EB/OL]. (2016-03-25)[2021-01-26]. https://www.bbc.com/news/technology-35902104.

[22] 杨庆峰.数据偏见是否可以消除?[J].自然辩证法研究,2019,35(8):109-113.

[23] O'NEIL C. Weapons of math destruction: how big data increases inequality and threatens democracy [M]. New York: Crown

Publisher, 2016.
[24] COUGHLAN S, SELLGREN K, BURNS J. A-levels: anger over 'unfair' results this year [EB/OL].(2020-08-13)[2021-01-26]. https://www.bbc.com/news/education-53759832.
[25] HARRIS E. Court vacates Long Island teacher's evaluation tied to test scores [EB/OL].(2016-05-10)[2021-01-26]. https://www.nytimes.com/2016/05/11/nyregion/court-vacates-long-island-teachers-evaluation-tied-to-student-test-scores.html.
[26] WALSH E, ISENBERG E. How does value added compare to student growth percentiles?[J]. Statistics and public policy, 2015, 2(1): 1-13.
[27] WHITE HOUSE. Big data: seizing opportunities, preserving values[R]. Washington: Executive Office of the President, 2014.
[28] ANGWIN J, Mattu S, LARSON J. The tiger mom tax: Asians are nearly twice as likely to get a higher price from Princeton review [EB/OL].(2015-09-01)[2021-01-26]. https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review.
[29] PARISER E. The filter bubble: what the Internet is hiding from you[M]. London: Penguin UK, 2011.
[30] FLAXMAN S, GOEL S, RAO J M. Filter bubbles, echo chambers, and online news consumption [J]. Public opinion quarterly, 2016, 80(S1): 298-320.
[31] PRAHALAD C K, RAMASWAMY V. Co-creation experiences: the next practice in value creation [J]. Journal of interactive marketing, 2004, 18(3): 5-14.
[32] 张林.智能算法推荐的意识形态风险及其治理[J].探索,2021(1):176-188.
[33] VORHIES W. Is model bias a threat to equal and fair treatment? maybe, maybe not[EB/OL]. (2018-06-05)[2021-01-26]. https://www.datasciencecentral.com/profiles/blogs/is-model-bias-a-threat-to-equal-and-fair-treatment-maybe-maybe-no.
[34] SOLOMON R C. The passions: the myth and nature of human emotion[M]. New York: Anchor, 1976.

Risk Analysis and Governance of Bias in Artificial Intelligence in Education

SHEN Yuan, WANG Qiong
(Graduate School of Education, Peking University, Beijing 100871)

[Abstract] In the context of advocating the inclusiveness and equity in education, education now faces the risk of bias raised by intelligent technologies characterized by individualization and self-adaptation. In the development of intelligent educational products, the potential design bias, data bias and algorithmic bias can expose the education sector to multiple risks, such as the hindrance of equity in educational activities, the increase of discrimination within the education system, the dissipation of the diversity in education and the narrowing of students' thoughts. Schools, technology companies, regulatory agencies and other organizations should work together to detect the risks of bias in advance and strengthen governance, including improving the educational understanding and diverse perspectives of research and development teams, and enabling teachers and students to become the co-creators of products, strengthening the process of bias detection and ethical assessment, and establishing systematic optimization mechanisms that can be explained, reviewed and corrected, carrying out AI literacy education for teachers and students, and establishing the scope and ethical norms of AI application in education, thus striving to achieve human-machine benign interaction and to create a diverse and inclusive intelligent education environment.

[Keywords] Artificial Intelligence in Education (AIED); Ethics of Artificial Intelligence; Educational Inclusiveness; Algorithmic Bias; Data Bias