算法公平：教育人工智能算法偏见的逻辑与治理

王佑镁　王　旦　王海洁　柳晨晨

（温州大学 大数据与智慧教育研究中心，浙江温州 325035）

［摘要］ 算法公平被视作人工智能领域的核心伦理问题。教育人工智能同样面临算法偏见、算法歧视等伦理风险。通过系统性回顾 2013—2023 年教育人工智能算法偏见的 57 篇文献，本研究发现，从总体架构看，教育人工智能算法偏见研究主要分概念性理论研究、教育场景应用研究与算法检测设计研究三类；从以算法偏见为核心厘清算法伦理风险、算法歧视、教育公平三层概念的逻辑层次看，研究样本在算法偏见的种类、成因和治理原则与方法方面存在共性和明确的指向。本研究最后以种类和成因为基础，提出未来教育算法偏见研究的五个方向，以期促进教育的算法公平，优化人工智能教育应用环境，推动教育人工智能生态系统的健康发展。

［关键词］ 教育人工智能；算法偏见；算法公平

［中图分类号］ G434　　［文献标识码］ A　　［文章编号］ 1007-2179(2023)05-0037-10

一、引言

教育人工智能指人工智能技术在教育领域的深化应用，即为教育创设高交互、泛在化与强感知的教与学环境，构建公平且有质量的教育人工智能系统。人工智能技术的快速迭代正逐步重塑教育生态。例如，ChatGPT 风靡海内外教育圈，成为互联网上最受关注的人工智能应用之一（王佑镁等，2023）。人工智能的教育应用以其个性化、定制化、智能化等优势可替代部分教育教学工作与管理，甚至教育决策（倪琴等，2022）。然而，这不可避免地会引发算法偏见、算法歧视等伦理风险，破坏算法公平，降低教育透明度，增添教育不确定性，导致教育不公平。

算法公平是人工智能领域的核心伦理目标，也是教育领域面临的重要伦理挑战，包括算法偏见、算法歧视、算法黑箱等。其中，算法偏见隶属算法伦理风险，指计算机执行计算或解决问题指令时，因初始算法、样本数据和歧视模仿等造成的思维处理惯性，导致人工智能系统运行出现偏向性的举措或选择。不同算法会产生不同类型的算法偏见（陈洪兵等，2019）。算法偏见主要来自于人类的偏见，以隐性偏见为主，但可识别与监督，尤其是显性偏见。教育领域算法偏见逐渐常态化，以性别偏见、种族偏见和地域差异等最为常见（Baker et al., 2021）。当下，人类主体作为潜在的算法开发者或使用者，

［收稿日期］2023-07-16　　［修回日期］2023-07-28　　［DOI 编码］10.13966/j.cnki.kfjyyj.2023.05.004

［基金项目］ 全国教育科学规划 2021 年度国家一般课题“教育领域人工智能应用的伦理风险与防范对策研究”（BCA210086）。

［作者简介］ 王佑镁，教授，博士生导师，温州大学大数据与智慧教育研究中心（wangyoumei@126.com）；王旦，王海洁，在读硕士，温州大学大数据与智慧教育研究中心（wdzzrs@163.com，whjzwx@163.com）；柳晨晨，博士，副教授，温州大学大数据与智慧教育研究中心。

［引用信息］ 王佑镁，王旦，王海洁，柳晨晨（2023）. 算法公平：教育人工智能算法偏见的逻辑与治理［J］. 开放教育研究，29(5)：37-46.

无意中会盲目跟从算法偏见，导致出现算法歧视（孟令宇，2022）。例如，教师使用人工智能技术未发现其存在的算法偏见，可能会对学生造成不良影响。鉴于教育领域的特殊性与重要性，未来的教育人工智能应用必须以人本主义为指导原则，重视教育数据与算法的伦理性，坚持透明且可解释原则，防止从算法偏见向算法歧视演变，及时纠偏教育人工智能算法伦理风险问题（曾海军等，2022）。2023年6月，中共中央办公厅等（2023）发布《关于构建优质均衡的基本公共教育服务体系的意见》提出，发展更高质量、更加公平的教育公共服务，是加快推进教育现代化并建设教育强国的重要方面。因此，消除数据和算法的偏见，减少教育不公平现象，维护个体权利，最终建立公平和机会平等为基础的教育体系，是教育人工智能应用的重要目标。本文以近十年教育人工智能算法偏见研究为样本，沿着“发文趋势→总体架构→逻辑层次→深层剖析→治理重心→未来方向”的思路展开探讨，以期为降低算法歧视的不良影响提供理论参考。

二、研究架构

（一）研究检索与筛选过程

本研究筛选2013年1月至2023年3月刊发的相关文献，首先以科学引文数据库（Web of Science）为基础来源，以“Artificial intelligence in education” or “AI in education” or “Education combined with artificial intelligence”与“algorithmic ethics” or “algorithmic bias” or “algorithmic discrimination” or “algorithmic fairness” or “algorithmic decision making” or “algorithmic risk”为关键词，搜索获得39篇英文文献。第一轮和第二轮筛选分别获得符合条件的文献22篇和9篇。本研究以相同的方式通过谷歌学术数据库搜索获得文献524篇。第一轮筛选获得32篇文献。排除会议、专著、社论材料等文献，第二轮筛选获得9篇文献。

中文文献以中国知网数据库为基础来源，将关键词“人工智能教育”或“教育人工智能”分别与“算法伦理”或“算法偏见”或“算法歧视”或“算法公平”或“算法决策”或“算法风险”任意一组关键词联合开展检索，排除相同文献后，共获得中文文献183篇。第一轮筛选后，符合条件的有48篇。排除会议、专著、社论材料等文献，第二轮筛选得到中文文献39篇。因此，本研究样本文献共57篇，其中英文文献18篇，中文文献39篇（检索方法与筛选流程见图1）。

（二）研究样本的分类与编码

本文将具有代表性的57篇文献作为分析对象，依据教育人工智能算法偏见的总体架构进行编码。为了宏观把握教育人工智能算法偏见的应用研究，57篇文献被分为三类：概念性理论研究、教育场景应用研究与算法检测设计研究，以明晰各文献的体系架构（见表1）。

1）概念性理论研究。本文将概念分为三类：上位概念、类似概念和影响效应概念。上位概念包括一般性伦理风险、算法伦理、教育偏见、算法风险等。一般性伦理风险指所有人工智能应用都需要面对的伦理风险，包括算法伦理风险（吴河江等，2020）。算法伦理风险包括算法偏见。其中，出现次数较多（4次及以上）的是算法决策和算法黑箱。类似概念表示与算法偏见相近却不同的概念，包括算法陷阱、算法技术规训的教育困境、歧视意识决策、刻板印象等。其中，出现次数最多的是算法歧视。影响效应概念表示因教育人工智能算法偏见而引发的影响与效应，包括权力与价值观、师生关系、教育理念与目标、教育决策质量等。其中，出现次数最多的是（4次及以上）权力与价值观、教育公平和社会影响。总体来说，影响效应概念被提及的次数最多，表明教育工作者较关心算法偏见带来的影响，关注人工智能对教育带来哪些伦理风险。

2）教育场景应用研究。本文将其分为两类：教育场景分类与软硬件。教育场景分类表示发生在不同学生状态、学生组织和学习关系的场景（袁凡等，2022），包括思想政治教育、教师教育、基础教育、精准教育等（赵磊磊等，2022）。其中，出现次数较多（4次及以上）的是思想政治教育和个性化教育。软硬件指应用于教育领域的算法监测与评估系统，包括人脸识别系统、自动评估系统、监督学习系统、虚拟现实环境等。例如，人脸识别技术用于监视学生的上课状态可能存在一定算法偏见（Akgun et al.，2021）。其中，出现次数最多（4次及以上）的是人脸识别系统。总体来说，教育工作者

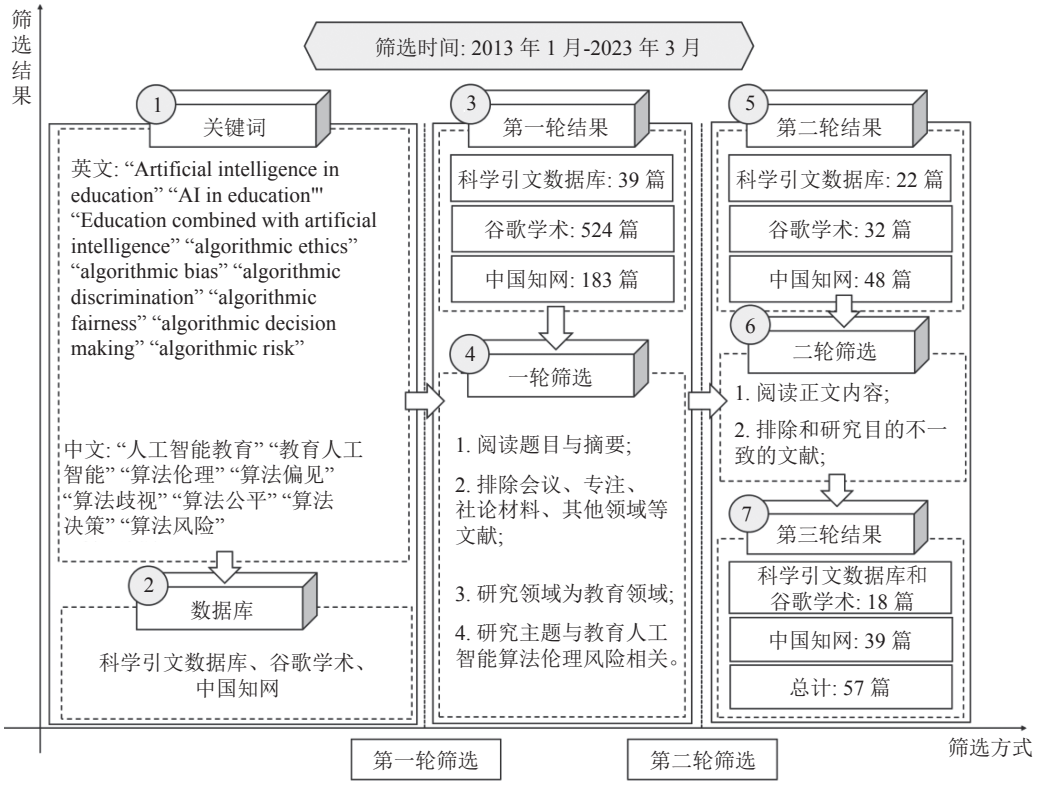

图 1　研究样本的检索方法与筛选流程

阐述算法偏见时，更偏向于从教育场景寻找解决算法偏见的办法。这可能是因为教育偏向人文社科类，而算法技术系统偏向理工科类。

3）算法检测设计研究。此类研究探讨通过算法监测工具，衡量技术应用于教育是否会发生算法偏见。例如，招生平台系统是否存在算法偏见，并提出需要对算法系统的利益相关者开展算法公平性、问责制、透明度和道德私领域的培训。国外这类研究已细化到研究某类算法的偏见，更有深度。

三、概念逻辑

比较教育人工智能算法偏见的研究总体架构可以发现，在概念性理论研究中，教育工作者常对算法偏见的相关概念，即上位概念、类似概念和影响效应概念之间的关系存疑，多数研究只简单描述三者之间的逻辑关系，未深挖其存在的关联。此外，研究样本多次提及未知偏见或隐性偏见。例如，罗斯皮利奥西（Rospigliosi, 2021）认为教育人工智能应用不仅包括显性偏见，还存在隐性偏见或未知的影响。这就需要以算法偏见为核心，厘清三类概念之间的逻辑关系，防范算法偏见衍生为算法歧视。

根据上文总结得出的 A1-C9 编码，本研究将上位概念、类似概念、影响效应概念作为三层逻辑，第一层与第二层的关系为包含，第二层与第三层的关系为作用，第一层以算法伦理风险为核心，第二层以算法偏见为核心，第三层以教育公平为核心，社会影响为次要核心（见图 2）。

第一层有三种关系，分别为交叉、原因与包含关系。1）交叉关系：①算法风险与算法伦理共同与算法伦理风险存在交叉关系；②算法技术规训的困境，包括数据依赖、算法偏见和强制的分析结果（孔苏，2021）；③教育偏见与算法伦理风险概念交叉的部分为算法偏见。2）原因关系：①算法决策与算法黑箱导致算法伦理风险的发生；②算法推荐导致用户陷入信息茧房，从而导致算法伦理风险的发生。3）包含关系：一般性伦理风险包括数据伦理风险和算法伦理风险（吴河江等，2020）。

第二层涉及三种关系，以线条表示，分别为递进、交叉与修饰关系。1）递进关系：算法偏见主要来自于开发者的偏见、数据的偏见和算法自身偏

表 1　研究样本的编码体系与统计

一级维度	二级维度	三级维度	编码	国内	占比(%)	国外	占比(%)	总占比
概念性理论研究	1.上位概念	一般性伦理风险	A1	1	1.7	0	0	25(43.9%)
		算法伦理	A2	2	3.5	0	0	
		教育偏见	A3	2	3.5	0	0	
		算法风险	A4	3	5.3	0	0	
		算法决策	A5	2	3.5	2	3.5	
		算法伦理风险	A6	3	5.3	0	0	
		算法推荐	A7	3	5.3	0	0	
		信息茧房	A8	1	1.7	0	0	
		算法黑箱	A9	4	7.0	1	1.7	
		算法技术规训的教育困境	A10	1	1.7	0	0	
	2.类似概念	算法陷阱	B1	1	1.7	0	0	21(36.8%)
		歧视意识决策	B2	0	0	1	1.7	
		刻板印象	B4	0	0	1	1.7	
		抑制剂	B5	0	0	1	1.7	
		算法歧视	B6	12	21.0	3	5.3	
		算法偏好	B7	1	1.7	1	1.7	
	3.作用效应概念	权力与价值观	C1	4	7.0	1	1.7	28(49.12%)
		师生关系	C2	2	3.5	0	0	
		教育理念与目标	C3	1	1.7	1	1.7	
		教育决策质量	C4	1	1.7	0	0	
		算法公平	C5	2	3.5	0	0	
		教育公平	C6	4	7.0	3	5.3	
		社会作用	C7	3	5.3	4	7.0	
		人的意志、行为与自主性	C8	1	1.7	0	0	
		数字鸿沟	C9	1	1.7	0	0	
教育场景应用研究	1.教育场景分类	思想政治教育	D1	5	8.8	0	0	18(31.6%)
		教师教育	D2	1	1.7	0	0	
		基础教育	D3	2	3.5	0	0	
		精准教育	D4	0	0	1	1.7	
		个性化教育	D5	3	5.3	3	5.3	
		教学评价	D6	2	3.5	1	1.7	
	2.软硬件	人脸识别	E1	2	3.5	2	3.5	9(15.8%)
		自动评估系统	E2	0	0	2	3.5	
		监督学习系统	E3	0	0	1	1.7	
		虚拟现实环境	E4	0	0	1	1.7	
		元宇宙学习应用	E5	1	1.7	0	0	
算法检测设计研究	1.检测招生平台中的算法偏见		F1	0	0	1	1.7	3(5.3%)
	2.针对性别和肤色，判断算法偏见		F2	0	0	2	3.5	

见，而算法歧视的责任主体主要是人。人类对算法偏见的盲从导致算法歧视(孟令宇，2022)。2)交叉关系：①算法决策出现算法歧视伦理风险的本质是人类固有的偏见和歧视意识决策(黎常等，2021)；②权利失范、关系失衡与情感异化间接导致算法陷阱的出现，也能加深算法偏见的影响(罗江华等，2022)；③刻板印象存有的算法偏见可描述为抑制剂。3)修饰关系：刻板印象和抑制剂均可用来修饰

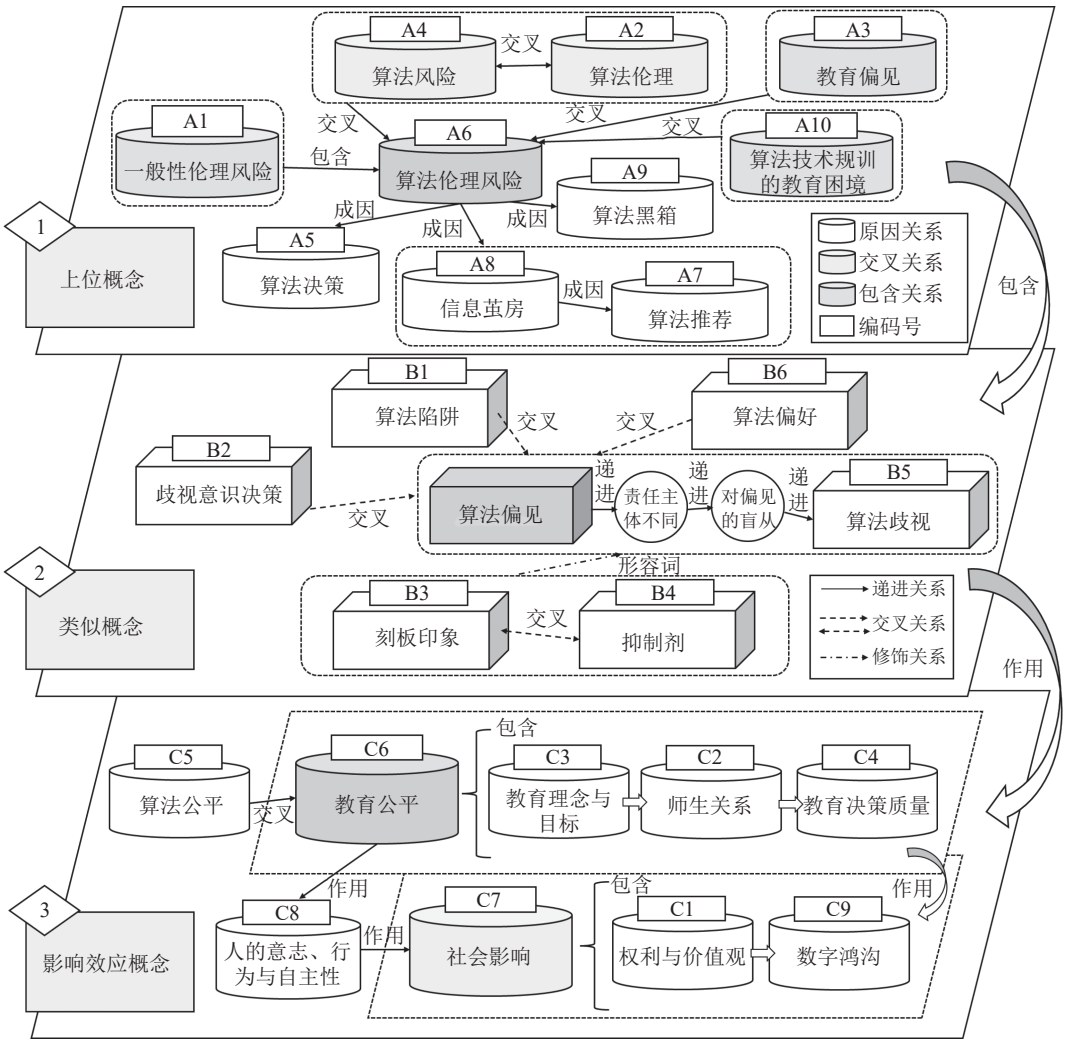

图2　三个层面的概念关系

算法偏见的描述方式。

第三层涉及三种关系，以不同的线条呈现，分别为包含、交叉与作用关系。1)交叉关系：算法公平性与教育公平性追求的最终目标——以人为本。2)包含关系：①教育公平涉及教育理念与目标、师生关系、教育决策质量；②社会影响包括权力与价值观遭到破坏，加剧数字鸿沟与差距。3)作用关系：①教育的公平性影响教育主体的意志、行为与自主性，从而对社会产生正面或负面影响；②教育的公平性直接对社会产生积极或消极的效应。

本文重点关注消极性算法偏见，探讨算法偏见与算法歧视之间的区别与关系，剖析算法伦理风险的组成要素，加深对教育界近期较为关注的算法推荐、算法黑箱、信息茧房等概念的理解。

1)算法偏见的定义。算法偏见是算法程序在信息的生产与分发过程中失去客观中立的立场，片面或与客观实际不符的信息、观念，影响公众对信息的客观全面认知(郭小平等，2019)。在教育领域，人工智能算法系统可能出现数据不平衡、男女教育资源不同、发达地区学生知识面广和学生价值观异化等影响教育公平和以人为本的教育目标的偏见性风险(王旦等，2021)。

2)算法偏见与算法歧视。算法偏见与算法歧视属于消极性名词，歧视强调一种行为，偏见强调一种认知状态。偏见可以是积极或消极的，甚至是中性的，消极的偏见可转化为歧视(孟令宇，2022)。在教育人工智能应用过程中，算法系统对学习者造成一系列算法偏见的伦理风险问题，教育工作者对其持有盲从与跟风的态度，导致教育算法歧视，破

坏了教育伦理的基本原则。

3)算法推荐、算法黑箱与信息茧房。算法推荐指利用生成合成类、个性化推送类、检索过滤类技术向用户提供信息(王思北等，2022)。在教育领域，算法推荐有可能产生一系列伦理问题、价值问题和虚假信息问题，进而使学习者陷入算法黑箱与信息茧房。例如，李伟(2021)认为通过算法推荐机制精准投放广告，扭曲了学习群体的消费观。郎捷等(2020)从信息茧房角度出发，提出学校教育应关注学生信息窄化、媒介素养低等问题，主动传播正能量，帮助学生群体树立正确的价值观。赵龙轩等(2022)以北京大学生为问卷调查对象，对其算法意识、算法态度及算法操纵行为之间的关系建构模型，为解决算法黑箱提供了路径。

四、偏见剖析

正文内容是文章的关键，比较分析正文内容有利于理解57篇文献的结构组成、特征和重点。本文提取研究样本中关于算法偏见的种类、成因与治理原则与方法的关键词，旨在从不同角度分析教育人工智能应用中算法偏见产生的影响，以便深入掌握研究样本的内容和分析对象的核心组成与特征(见图3)(王佑镁等，2021)。

第一，教育人工智能算法偏见的种类。总体来看，提及次数最多的是性别偏见(14次)和种族偏见(9次)，特殊群体、家庭与收入、政治和地域差异的偏见为6次及以下，政治偏见提及次数最少(3次)。其中，国外提及最多的是性别偏见(11次)和种族偏见(8次)。出现这一差别的原因可能是国外教育人工智能算法偏见研究较关注种族和性别偏见。

第二，教育人工智能算法偏见的成因。总体来看，国内外提及次数最多的是人类的不确定性(15次)和设计责任人的无意偏见(14次)。多元主体的博弈(4次)和无监测机制(3次)提及较少。原因可能是教育领域人工智能应用占主导的依然是教育主体和系统背后的设计开发者，他们对诱发算法偏见影响较大。此外，历史大数据的偏差(9次)与系统自带的隐性偏见(8次)亦不容忽视。算法模型基于历史数据训练而成，会使人工智能算法系统自带隐性偏见。这些也是教育工作者不容小觑的成因要素。

第三，教育人工智能算法偏见的治理原则与方法。在治理原则方面，教育工作者重视提出解决措施。其中，可透明性(17次)、问责制(10次)、公平正义(12次)与系统化监管(11次)四条原则被提及次数较多，多元化和包容性(5次)和以人为本(4次)被提及次数较少。国内学者提及较多的治理原则是可透明性和系统化监管。例如，谭维智(2019)提出保持算法和计算模型的透明化，建立算法教育应用风险评估和监管机制。国外学者提及较多的是可透明性和公平正义。例如，泽德(Zeide，2019)指出当算法的评估建议和结果与教育工作者的专业判断不一致时，学校需制定清晰的标准，遵从教育的公平正义原则。在治理方法方面，提及次数较多的十项解决路径可分两类：多元主体协助解决与教育主体主导治理。其中，多元主体主要为积极提高数据素养、责任意识和态度，并将技术赋能算法治理，与人文情感巧妙结合。教育主体主导治理主要包括让使用者自我监控问题所在、教师参与决策与预先测试、开发人员嵌入道德规则、教育管理者制定治理协议和风险评估管理的规则等。

五、偏见治理

根据上文总结的教育人工智能算法偏见的种类、成因与治理原则和方法，本研究以种类和成因为基础，归纳出教育人工智能算法偏见治理的重点(见图4)：

(一)从准确到精准，改进历史数据的收集方式

教育人工智能应用应注意教育产品中早已嵌入的无意或未知偏见。大数据本身带有成见，其片面性易产生算法偏见风险，造成算法歧视(靖东阁，2021)。算法教育一般依赖于教育大数据，而教育大数据常从现实生活抽取而来，不可避免地带有教育发展过程中固有的属性与特征。一旦输入的教育数据有偏差，那么算法系统的推算结果就可能带有偏见。历史数据的偏见、数据采样不充分和数据集特征选取不当等都可能为教育人工智能应用带来局限。多数研究提到算法偏见很大程度上取决于历史数据的透明度，但教育人工智能产品中算法系统自带的算法偏见是造成伦理风险的重要成因。阿克古恩等(Akgun et al.，2021)认为，尽管算

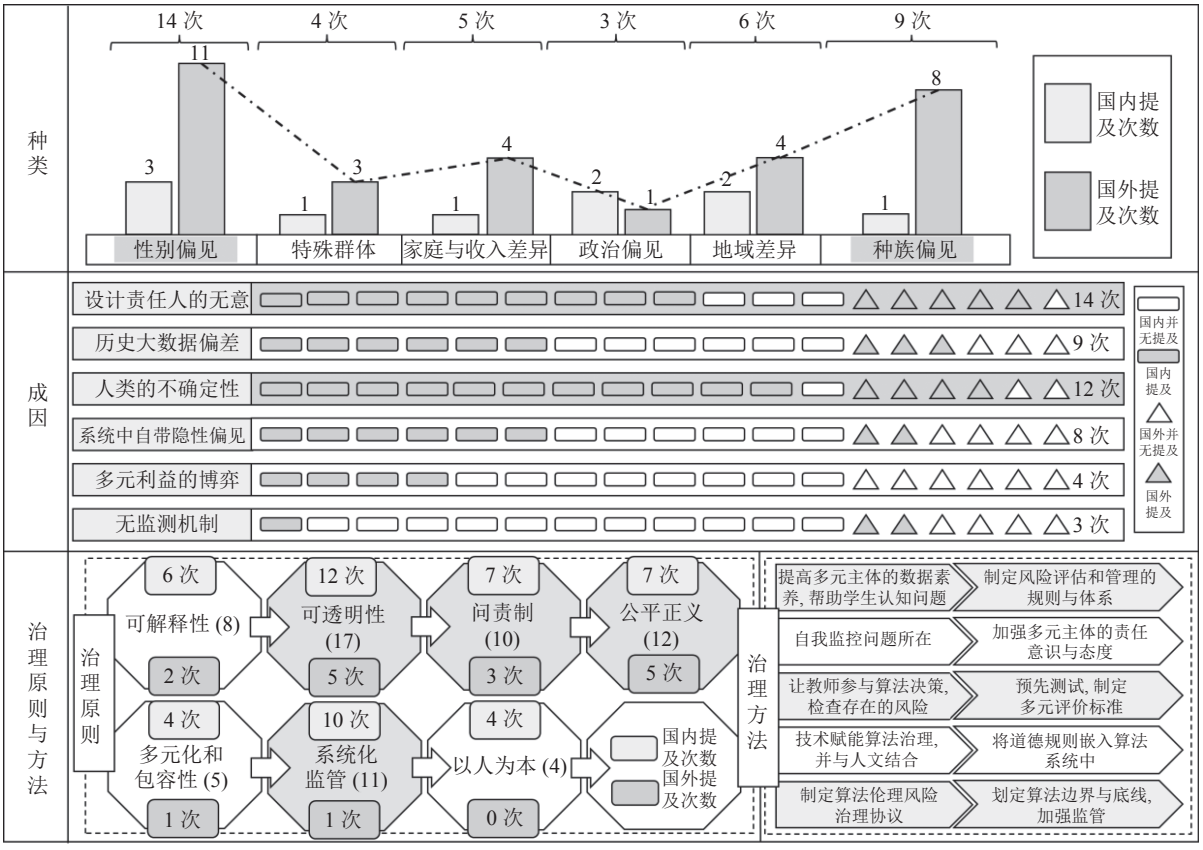

图 3　算法偏见的种类、成因、治理原则与方法的分析框架

法模型没有明显地嵌入偏见，但依然能在教育人工智能平台发现有关性别和种族的偏见。因此，教育数据的收集方式、代表性和质量等需要改进与升级，从准确性向精准性收集蜕变（肖凤翔等，2020）。例如，学生利用人工智能教育产品学习数学，平台算法可能倾向于向男生推荐有挑战性的问题，向女生推荐较简单的问题，那么这类教育平台就需要寻求更多样化的数据，并对算法进行反偏见训练。只有正确衡量与矫正历史数据，才能切断算法偏见的源头。

（二）从规范到系统，制定评估与监管标准体系

可持续的教育人工智能应用离不开教育风险评估与监管人才队伍的培养。有研究者提出人工智能算法决策的数据处理方式均由“人”创建，一定带有个体主观的隐性偏见（赵磊磊等，2022）。当主观的算法偏见渗透于教师教育管理过程，教师的决策质量就会下滑，导致学生被不公平对待。这就需要制定有效的人工智能与教育伦理风险评估与监管标准体系，以人本主义为指导原则，确保教育数据与算法合乎伦理，具备透明性、可解释性和可纠正性，促进教育人工智能应用高水平安全和高质量发展的良性互动。以“歧视感知数据挖掘技术”为代表的技术，能通过“算法”识别程序判断算法决策是否存在算法伦理风险，这是实现教育人工智能应用算法公平的有效方式之一（潘芳芳，2021）。

（三）从个体到团队，多元主体协作共治

多元主体协作共治指开发者、企业、教育决策者、教育者、学习者和其他利益相关者从独立参与转为协作共建与共治。这就需要提高全员算法素养，开展系统的科技伦理教育，承认个体差异性和权利，坚持公平公正和公开透明的算法机制，保证学生的合法利益和尊严不受侵犯（冯建军，2022）。其中，开发人员应在教育人工智能产品的算法系统中嵌入道德伦理规则，增强伦理意识，坚持正确的价值观，认真履行伦理培训责任，确保智能算法始终坚守“以人为本”；企业应提前预估发生算法偏见的负面影响，保持智能算法的透明度和可解释性，与开发人员合作为学生打造良性的教育环境；教育

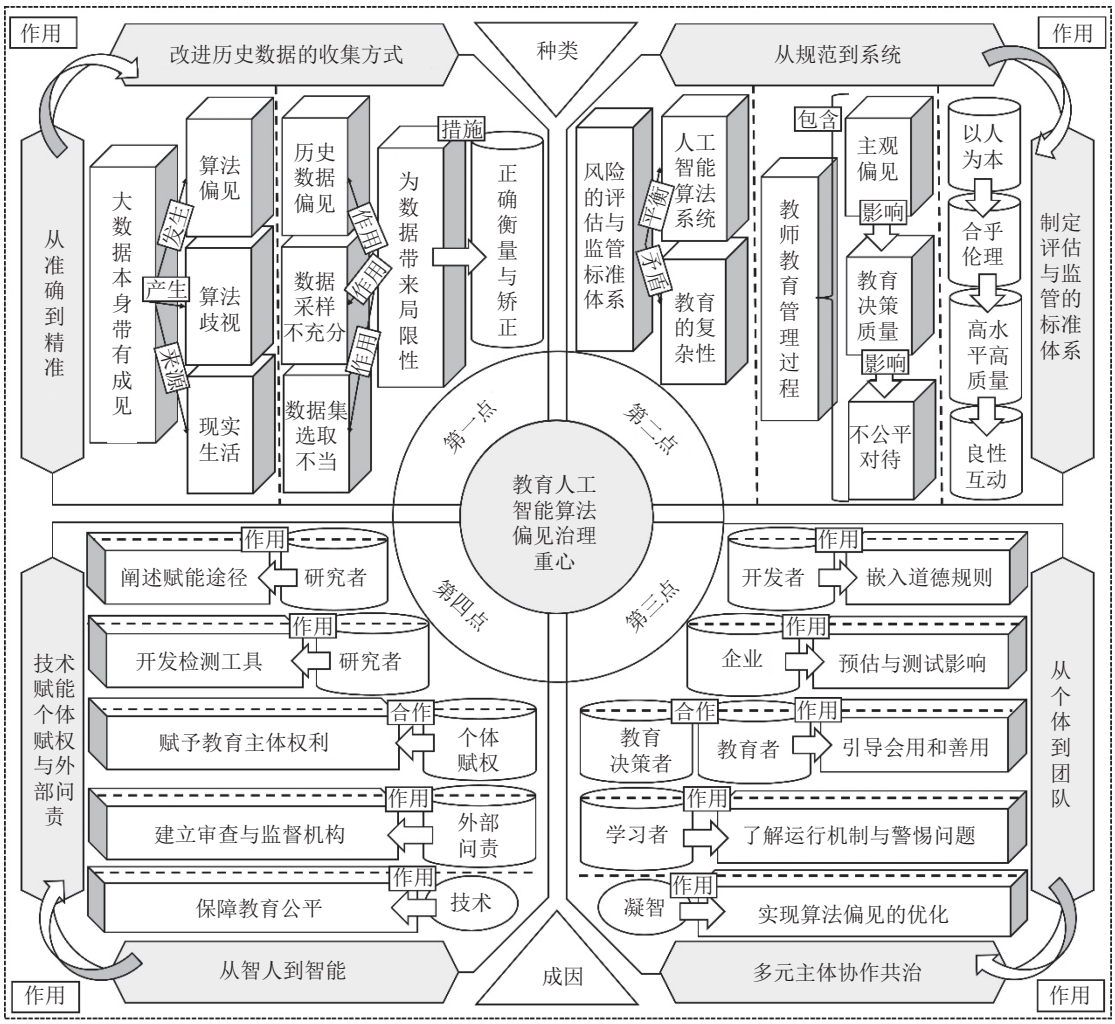

图 4　治理算法偏见的四点路径

工作者应引导学生会用和善用智能技术，参与算法决策，检查与测试自动化监考软件、情绪识别系统等产品是否存在偏见，并提出有效解决算法偏见的建议，将学生的主动性和发展性作为智慧教学的落脚点；学校要为学习者逐步开设人工智能算法课程，帮助学习者了解算法运行机制，知悉算法偏见造成的社会影响，警惕算法偏见、算法歧视、算法黑箱等不易发现的问题（向勇，2021）。

（四）从智人到智能，技术增强个体赋权与外部问责

随着人工智能技术的不断发展，算法伦理风险的治理方式正从人治向智能化转变。有研究者详细阐述了通过技术赋能治理算法伦理风险的途径，并开发算法偏见的监测工具验证教育人工智能应用中算法偏见的发生（孔苏，2021；Sha et al.，2022）。例如，在教育人工智能应用中，算法往往根据历史数据作出可能带有偏见或歧视的决策。这通过算法检测工具可减少人类重复性审查工作，实现智能化的修正，保证算法决策的公平。技术赋能包括教育主体的赋权和外部问责机制，即采用技术纠正偏见的方法为教育领域人工智能的应用提供平等的决策。在个体赋权方面，教师、家长或学生等应了解和参与算法的运作过程，提供反馈和建议（杨利华等，2023）。在此过程中，利用个性化算法可为教育主体定制符合需求的算法课程与学习任务，采用可解释性人工智能技术帮助其理解算法决策。外部问责可借鉴发达国家的经验，建立专门机构审查与监督教育人工智能的应用。各地教育部门应互相支持与合作完善问责机制，对存在算法伦理风险的教育产品和服务立即终止生产与使用，做到对相

关企业和责任人的严肃追责和问责管理，有效落实实名制审查与监督机制（孙伟平等，2020）。

六、未来研究方向

教育人工智能算法偏见研究需重点关注以下方面：

1）在无法完全消除算法偏见时，教育领域应重点研究由此引发的算法歧视风险，减少人类主观意识、情感和行为引发的风险，以及算法歧视对学生的风险影响，形成有效的治理路径，确保教育的算法应用公平和平等。

2）关注智能化教育的算法决策、算法推荐、算法黑箱和信息茧房，理论联系实际，产出更多有价值的研究，形成治理规划，不断改进智慧化教与学环境，确保算法发挥更加积极的作用。

3）当前教育人工智能算法偏见研究仅关注少量教育场景，未来仍需探索职业教育、职前教育等教育场景的算法伦理风险，综合研究多重场景的算法伦理风险。

4）关注生成式人工智能教育产品可能存在的偏见问题。鉴于大模型技术水平落后，未来我国应大力开发算法偏见检测工具，引进并学习发达国家技术经验，倒逼国内的技术进步，从而更好地处理教育场景的算法偏见及类似的伦理风险，提升治理与防范水平，创造更加包容和公正的智能教育环境（姚树洁等，2021）。

5）未来的治理重点包括系统化地制定教育人工智能算法偏见的风险评估协议、风险监测机制和规范化治理原则。例如，国家人工智能治理专业委员会发布的《新一代人工智能伦理规范》第十三条，特别强调避免算法偏见和歧视，强化伦理审查。这一指引为国内实现教育人工智能系统的公平性、普惠性和非歧视性提供了方向。未来，我们需要建设多元共建与共治的防范体系，与利益相关者形成共识，确保教育人工智能的合理应用。

［参考文献］

[1] Akgun, S., & Greenhow, C.（2021）. Artificial intelligence in education: Addressing ethical challenges in K-12 settings[J]. AI and Ethics（2）: 431-440.

[2] Baker, R. S., & Hawn, A.（2021）. Algorithmic bias in education[J]. International Journal of Artificial Intelligence in Education, (32): 1052-1092.

[3] 曾海军，张钰，苗苗（2022）. 确保人工智能服务共同利益，促进教育系统变革——《人工智能与教育：政策制定者指南》解读 [J]. 中国电化教育，427（8）：1-8.

[4] 陈洪兵，陈禹衡（2019）. 刑法领域的新挑战：人工智能的算法偏见 [J]. 广西大学学报（哲学社会科学版），41（5）：85-93.

[5] 冯建军（2022）. 网络公民教育：智能时代道德教育的新要求 [J]. 伦理学研究，119（3）：1-9.

[6] 郭小平，秦艺轩（2019）. 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），41（9）：19-24.

[7] 靖东阁（2021）. 人工智能时代教育研究的计算主义及超越 [J]. 电化教育研究，42（2）：18-24.

[8] 孔苏（2021）. 智能教育的算法技术规训困境与出路 [J]. 电化教育研究，42（12）：36-40+54.

[9] 郎捷，王军（2020）. “信息茧房”对大学生思想政治教育的挑战及应对分析 [J]. 学校党建与思想教育，635（20）：13-15.

[10] 黎常，金杨华（2021）. 科技伦理视角下的人工智能研究 [J]. 科研管理，42（8）：9-16.

[11] 李伟（2021）. 消费主义对大学生社会心态的影响及其应对 [J]. 河南社会科学，29（11）：10-18.

[12] 罗江华，王琳，刘璐（2022）. 人工智能赋能课堂反馈的伦理困境及风险化解 [J]. 现代远程教育研究，34（2）：29-36.

[13] 孟令宇（2022）. 从算法偏见到算法歧视：算法歧视的责任问题探究 [J]. 东北大学学报（社会科学版），24（1）：1-9.

[14] 倪琴，刘志，郝煜佳，贺樑（2022）. 智能教育场景下的算法歧视：潜在风险、成因剖析与治理策略 [J]. 中国电化教育，431（12）：93-100.

[15] 潘芳芳（2021）. 算法歧视的民事责任形态 [J]. 华东政法大学学报，24（5）：55-68.

[16] Rospigliosi, P. A.（2021）. The risk of algorithmic injustice for interactive learning environments[J]. Interactive Learning Environments, 29(4): 523-526.

[17] Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G.（2022）. Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education[J]. IEEE Transactions on Learning Technologies, 15(4): 481-492.

[18] 孙伟平，伏志强（2020）. 智能时代教育公正问题探析 [J]. 湖北大学学报（哲学社会科学版），47（4）：25-31.

[19] 谭维智（2019）. 人工智能教育应用的算法风险 [J]. 开放教育研究，25（6）：20-30.

[20] 王旦，张熙，侯浩翔（2021）. 智能时代的教育伦理风险及应然向度 [J]. 教育研究与实验，201（4）：34-39+96.

[21] 王思北，阳娜，周琳，颜之宏（2022）. 大数据“杀熟”不能再“杀”了，算法推荐不能乱“推”了 [N]. 新华每日电讯，2022-01-07(7).

[22] 王佑镁，宛平，柳晨晨（2021）. 培养负责任的数字公民——国际数字公民教育政策文本的多维比较 [J]. 比较教育研究，43（3）：8-14+23.

[23] 王佑镁，王旦，梁炜怡等（2023）. “阿拉丁神灯”还是“潘多拉魔盒”：ChatGPT 教育应用的潜能与风险 [J]. 现代远程教育研究，35（2）：48-56.

[24] 吴河江，涂艳国，谭轵纱（2020）. 人工智能时代的教育风险及其规避 [J]. 现代教育技术，30（4）：18-24.

[25] 向勇. 算法推荐时代高校思想政治理论课的创新研究 [J]. 社会科学，2021，496（12）：70-80.

[26] 肖凤翔，张双志（2020）. 算法教育治理：技术逻辑、风险挑战与公共政策 [J]. 中国电化教育，396（1）：76-84.

[27] 杨利华，苏泽祺（2023）. 智能社会中算法治理的法律控制研究 [J]. 大理大学学报，8（1）：101-109.

[28] 姚树洁，房景（2021）. 科技创新推动“双循环”新格局发展的理论及战略对策 [J]. 东北师大学报（哲学社会科学版）（3）：39-51.

[29] 袁凡，陈卫东，徐铦忆等（2022）. 场景赋能：场景化设计及其教育应用展望——兼论元宇宙时代全场景学习的实现机制 [J]. 远程教育杂志，40（1）：15-25.

[30] Zeide, E.（2019）. Artificial intelligence in higher education: Applications, promise and perils, and ethical questions[J]. Educause Review, 54(3).

[31] 赵磊磊，张黎，代蕊华等（2022）. 人工智能赋能教师教育：基本逻辑与实践路向 [J]. 中国教育学刊，350（6）：14-21.

[32] 赵龙轩，林聪（2022）. “黑箱”中的青年：大学生群体的算法意识、算法态度与算法操纵 [J]. 中国青年研究，317（7）：20-30.

[33] 中国中央办公厅，国务院办公厅（2023）. 关于构建优质均衡的基本公共教育服务体系的意见 [OL]. [2023-08-26]. https://www.gov.cn/gongbao/2023/issue_10546/202306/content_6888957.html.

（编辑：魏志慧）

Algorithmic Fairness: Logical Levels and Governance Focus of Algorithmic Bias in Educational Artificial Intelligence

WANG Youmei, WANG Dan, WANG Haijie & LIU Chenchen

（*Research Center for Big Data and Smart Education*, *Wenzhou University*, *Wenzhou* 325035, *China*）

Abstract: *The issue of algorithmic fairness is seen as an ethical issue at the heart of the field of artificial intelligence. Educational AI also faces ethical risks such as algorithmic bias and algorithmic discrimination. A systematic review of* 57 *papers on algorithmic bias in educational AI from* 2013 *to* 2023 *found the research on algorithmic bias in educational AI is mainly divided into three categories*: *Conceptual theoretical research*, *research on the application of educational scenarios*, *and research on the design of algorithmic detection. The review also found that from the logical hierarchy of clarifying the ethical risks of algorithmic ethics*, *algorithmic discrimination*, *and educational fairness were the core of algorithmic bias. In addition*, *the review discovered that from the logical hierarchy of the three conceptual layers*, *the research samples have commonalities and clear directions in terms of the types*, *causes*, *and governance principles and methods of algorithmic bias. Based on the review results*, *the article proposes five directions for the development of future research on algorithmic bias in the field of education*: *Slowing down the evolution from bias to discrimination*, *promoting algorithmic fairness in education*, *optimizing the environment of AI educational applications*, *and promoting the healthy development of the educational AI ecosystem.*

Key words: *educational artificial intelligence*; *algorithm bias*; *algorithmic fairness*