DOI:10.16382/j.cnki.1000-5560.2023.03.006

教育数字化转型的核心技术引擎：可信教育人工智能*

江 波 丁莹雯 魏雨昂

（华东师范大学教育信息技术学系/上海智能教育研究院，上海 200062）

摘 要：教育数字化转型旨在通过数字技术实现教育教学流程再造和提质增效。一方面，以人工智能为代表的数字技术作为教育数字化转型的技术引擎，驱动教育数字化转型持续深入。另一方面，人工智能技术的“黑箱”本质引起了人机信任危机，存在违背教育中的公平、责任、透明、伦理等基本约束的风险，阻碍了教育数字化转型。本研究梳理了人工智能助力教育数字化转型过程中所引发和加剧的四大治理难题，剖析了可信教育人工智能的理论研究和发展现状，提出了可信教育人工智能的基本框架，总结了可信教育人工智能为教育数字化转型带来的机遇，提出了可信教育人工智能促进教育数字化转型的发展建议。研究建议出台可信教育人工智能的相关标准和法规，将技术可信度纳入教育数字化建设评价体系中，深入推进教育数据治理，提升教育从业者的数字素养。

关键词：教育数字化转型；可信人工智能；可信教育人工智能；教育治理

一、引言

数字技术的指数级发展速度远超于当前社会、政治、经济层面的适应能力（Hanna，2018），各行各业都在积极推进和落实数字化战略。数字化转型是使用人工智能、自动化、5G 等数字技术，利用数据实现工作的智能决策和实时响应[①]，是利用技术从根本上提高组织的绩效或者影响力（Brence & Mauhart，2019），是在信息技术应用不断创新和数据资源持续增长的双重作用下对相关领域的变革与重塑（翟云等，2021）。教育是数字化转型的重要领域。UNESCO（2022）认为数字化学习和转型是促进教育改革的关键杠杆。祝智庭等（2022a）将教育数字化转型定义为将数字技术整合到教育领域的各个层面，推动教育全方位创新与变革，实现包容有序、开放持续的良好教育生态。它通过教育数据驱动，数字技术赋能，实现教育教学的提质增效和传统教育教学流程的优化与重塑。

从技术变革教育的角度来看，我国教育发展大致经历了电气化、信息化 1.0、信息化 2.0 三个主要阶段（黄荣怀，2022）。信息化 2.0 阶段的一个显著特征是教育数字化转型，而在诸多数字化技术中，人工智能是能够真正实现提质增效的核心技术。无论是《教育信息化 2.0 行动计划》，还是《教育现代化 2035》，都强调充分利用人工智能技术助力教育教学的改革发展。以人工智能为代表的新一代数字技术正对传统教育生态、教育环境、教学方式、教育治理产生革命性影响，已然成为教育数字化转型的核心技术。近年来，各类智能教育平台纷纷将人工智能作为核心技术，在智能辅导、微格教学、自适应学习、沉浸学习、自动测评、课堂评价、数据决策、教育治理等多方面推动了教育的革新与发展（杨晓哲，任友群，2021）。人工智能技术助力教育的革新主要体现在教育教学流程再造、知识供给形态变

* 基金项目：国家自然科学基金项目“面向图形化编程的项目式学习的自动化评价研究及应用”（61977058）；上海市科技创新行动计划“人工智能”专项项目“教育数据治理与智能教育大脑关键技术研究与典型应用”（20511101600）。

革、教育评价模式优化和教育管理形式创新等四个方面（刘三女牙，2022）。

然而，人工智能的技术本质决定了它天生具有偏见和风险。微软将人工智能带来的挑战归结为公平、责任、透明、伦理（Fairness, Accountability, Transparency, Ethics, FATE）这四大核心问题[②]。教育的对象是人，人工智能教育应用的前提是平等、公平和公正地使用（张坤颖，张家年，2017）。人工智能的“黑箱”决策机制有违以人为中心的决策理念，因此难以被教育利益相关者接纳，而教育教学决策科学与否直接影响了教学效益（魏亚丽，张亮，2022）。人工智能的强大的预测能力在赋能教育教学决策的同时，也会放大教育决策的负面影响，从而违背了教育的公平、责任、透明和伦理等硬约束。因此，UNESCO 发布的报告（Duggan & Corporation，2020）和建议书（UNESCO，2021）均强调透明度、可监督性和可解释性对教育人工智能的重要性。由于教育决策的时效性和特殊性，教育人工智能应用需要在设计和开发阶段就实现透明性（Transparence）和可解释性（Explainability），而不能事后弥补。教育由复杂成分构成，它既是服务业，也是社会文化事业，还是极其复杂的社会现象（祝智庭等，2018）。教育治理问题高度复杂，提升教育人工智能模型的透明度和可解释性，不是一个简单的技术问题，更需要从个人、部门、社会等主体对教育人工智能的监督机制、隐私保护、伦理法规、公平决策等方面进行约束，从而构建起兼顾性能、可解释性、公平伦理、问责机制和透明度的可信教育人工智能体系。

二、人工智能赋能教育数字化转型的治理难题

本节将具体阐述人工智能赋能教育数字化转型带来的治理问题及引发的负面影响。

（一）人工智能加剧了教育数字化转型中的公平问题

教育公平一直是教育变革中的重点话题。平等和公平是教育公平中两个常见且易混淆的概念。Kizilcec 和 Lee（2020）总结了技术创新对教育公平的影响（如图 1 所示）。平等是通过技术创新活动给不同学生群体在整个教育过程提供同等的教育机会和资源，由于不同学生群体获取资源的途径及学习能力存在差距，导致群体间的差距保持不变或者扩大；而公平是通过技术创新活动将教育资源和机会向边缘弱势群体倾斜，从而缩小群体间差距。在人工智能赋能教育数字化转型过程中，可能会加剧教育不公平问题。在算法层面，人工智能算法中的数据样本不均衡和数据标签歧视，使得算法推送产生偏见和歧视，容易将学习者置于“信息茧房”，推送给学习者较为单一的学习内容，忽视其学习和发展的灵活性和多样性。在教育资源层面，应用对象往往向数字化水平高的地域倾斜，不同地区的学生对互联网等数字技术的应用存在差距，导致不同学生获得教育资源不等。在应用设计层面，产品开发者的逐利特性使其在产品设计阶段忽略特殊人群的需求，智能产品的包容性低、智能产品设备要求高和迭代快，给边缘群体带来困扰，甚至使其排斥、厌恶智能产品，“数字鸿沟”将随之拉大，最终加剧教育的不公平。

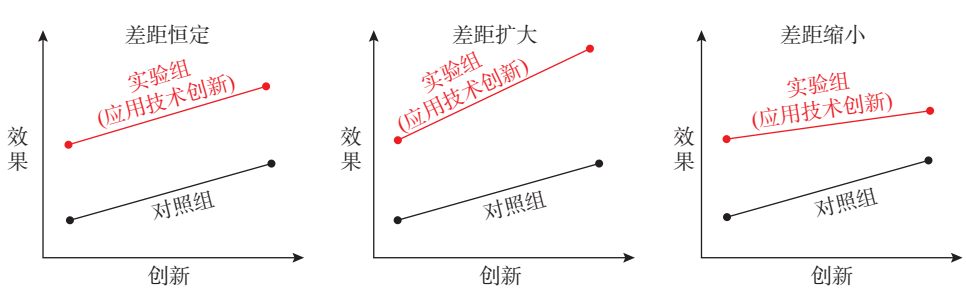

图 1　技术创新对教育公平的影响 (Kizilcec & Lee，2020)

（二）人工智能加剧了教育数字化转型中的责任问题

随人工智能算法的功能性增强、复杂性增加，其自主性增高，致使算法可解释性降低。当算法无法被观测和解释时，其责任分配和归属也就无从谈起（刘艳红，2022）。教育人工智能的责任问题是

指,谁来为人工智能带来的负面影响或具有争议的后果承担责任。在数据管理层面,由于在数据的采集、分析、处理和存储等过程中都有可能发生数据泄露、窃取和滥用的不正当行为,导致学习者个人数据泄露问责困难(胡姣等,2022)。在数据处理及共享层面,在数据共享和解释时可能对个别学生带上种族、地位等有色标签,使其合法权利受到威胁,此时责任主体无法明确(于聪,刘飞,2022)。在教育人工智能产品应用层面,教育人工智能产品可能会导致学习的失败,人工智能算法作为其中的责任主体,由于其过程不透明、内部算法复杂、决策"黑箱",很难确定系统的决策机制是如何运行的,也很难对其进行追溯和解释。不具有透明度和可解释性的人工智能,无法明确事故的发生因何而起,那么教育人工智能的问责机制就无法实现。

(三)人工智能加剧了教育数字化转型中的透明问题

教育人工智能的透明性是为保障教育利益相关者的知情权,公开关于教育人工智能系统的数据收集和管理框架、数据标签和清洗方法、影响特定预测或决策的算法等信息,可以增进教育利益相关者对系统的信任(苗逢春,2022)。教育人工智能的透明性体现在两个方面:一是模型的每个决策单元以及决策单元之间的可访问、可观测的程度;二是系统能够以利益相关者可理解的方式披露信息,如技术构成、行为逻辑、使用和维护方法及输出的预期结果等。教育场景的特殊性一定程度上约束了教育人工智能的透明机制的建立。一是技术的限制,我国教育利益相关者的数字素养整体有待提高,而当前的可解释性技术尚未达到"通俗"水平,导致利益相关者无法理解技术决策的过程。二是法律法规的限制,数据透明存在损害利益相关者个人隐私的潜在风险,相应的涉及个人隐私的数据标签可能未经用户知情允许而被公开,从而诱发信息窃取泄露的违法行为。

(四)人工智能加剧了教育数字化转型中的伦理问题

人工智能赋能教育数字化转型中面临的伦理问题包括利益相关者伦理问题、技术伦理问题和社会伦理问题(托雷•霍尔等,2022;苗逢春,2022)。在教育人工智能技术的设计过程中,带有"偏见基因"的数据和算法引发教育偏见、歧视和个人隐私泄露等问题。师生面临一系列伦理困境,如教师的角色定位、该以何种伦理准则保护学生个人隐私和信息安全等。学生作为教育人工智能产品最直接的受益者,受到智能学习服务的有效性因人而异,在学习者大脑发育的关键期接受不适宜的学习服务将导致其生理和心理、社会交往习惯和认知能力等逆向发展。例如,过度使用电子设备对学习者视力、听力等生理结构造成的不可逆影响及剥夺学习者与同伴、教师和家长的交流时间,从而造成社会交往能力缺失(荆敏菊,2015)。教育数字化转型中由于数据、算法及其决策过程导致技术伦理问题,从而引发教育不公的社会伦理问题。此外,教育人工智能技术中存在缺乏可解释性和透明度的技术伦理问题,导致责任主体无法被界定,从而引发问责难的社会伦理。

三、可信教育人工智能框架构建

综上所述,人工智能的教育应用在促进教育数字化转型的同时,也在不同程度上加剧了教育教学中的公平、责任、透明和伦理等问题。归根结底,人工智能赋能教育数字化转型而引发和加剧的治理问题是人工智能技术引起的。因此,如果能从技术源头实现透明性和可解释性,则可最大程度降低人工智能赋能教育数字化转型带来的负面影响。

(一)可信人工智能

信任是建立合作的基础,人与技术之间同样需要通过信任建立人机协同的合作关系。社会–技术系统中,人与技术的发展是相辅相成的,人类信任、接纳技术以减少劳动成本,发挥技术价值,进而促进技术革新。构建可信人工智能(Trustworthy AI, TAI)是解决信任问题的必要条件。Liu 等人(2022)分别从技术、用户、社会对 TAI 的原则进行总结:从技术本身考虑,TAI 需要高准确度、鲁棒性和可解释性;从用户出发,TAI 需要具备可靠可用性、安全隐私及自主性;从社会的角度,TAI 需要顾及法律、道德、公平和环保问题。如图 2 所示,欧盟在《可信赖的人工智能伦理准则》(European Commission, 2019)

中提出了 TAI 的框架，包括尊重人的自主性、防止伤害、可解释和公平等 4 个伦理原则，人类代理和人类监督、技术健壮性和安全性、隐私和数据管理、透明度、多样性、社会和环境福祉、问责等 7 个要素，以及技术方法和非技术方法等 2 类方法。可信人工智能的原则贯穿于整个人工智能系统的生命周期，并进行不断的评估反馈、迭代优化。IEEE 在 2022 年技术预测中将 TAI 确定为随后几年的领先新兴领域之一（IEEE Computer Society，2022）。

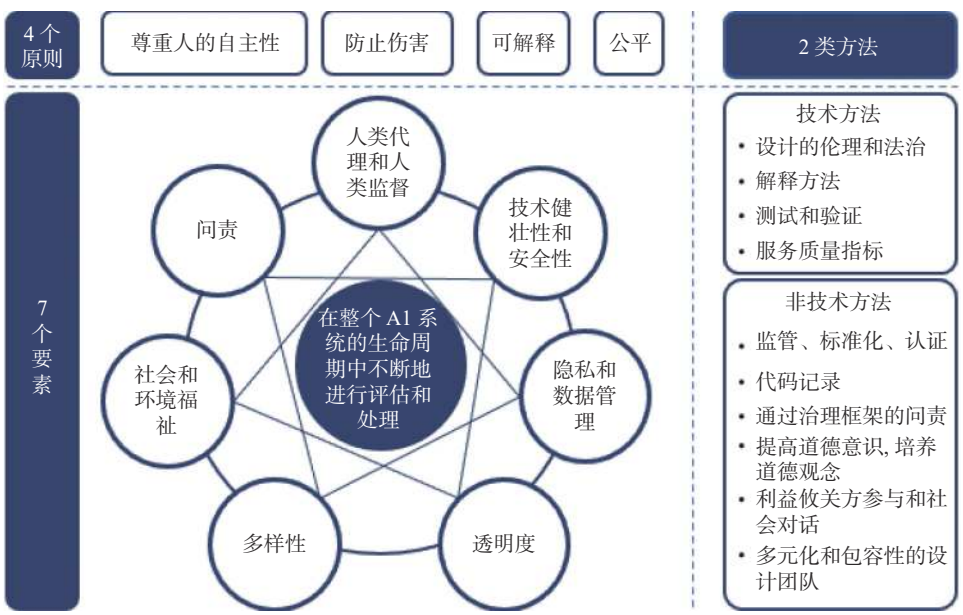

图 2　欧盟《可信赖的人工智能伦理准则》

（二）可信教育人工智能的基本框架

当前，可信人工智能已经上升到各国人工智能发展的重大战略。在教育领域，利益相关者尤为重视教学决策中的公平、责任、透明和伦理问题，当技术介入其中，算法“黑箱”、算法偏见等一系列问题导致教育过程变得机械化和形式化，教育参与者的主动权被削弱，教育过程与教育主体逐渐剥离（冯永刚，赵丹丹，2022）。在人工智能赋能教育数字化转型的背景之下，技术的可信度将会影响转型的“质”和“效”。可信人工智能的提出能在一定程度上缓解技术赋能教育数字化转型引发和加剧的治理难题。众多国际组织和学术团体尝试构建可信人工智能的通用框架与指南（中国通信院，京东探索研究院，2021; European Commission, 2019; Trump, 2020），但教育领域的专业性约束了通用指南在应用场景、作用对象和实现细节上的针对性。因此，本研究基于可信人工智能伦理准则的七大要素，结合教育场景的特殊性，提出了可信教育人工智能（Trustworthy Artificial Intelligence in Education, TAIE）框架，回答教育利益相关者在设计、开发和运用教育人工智能技术的过程中为何需要可信、何以实现可信。进一步，基于社会–技术系统理论（Geels, 2002），从技术、社会和人三个角度来阐述教育场景下可信人工智能的基本特征和要求。

社会–技术系统理论认为，技术系统的变革是社会经济、组织制度和社会文化等多种因素作用的结果（Geels, 2002）。技术创新促进社会发展是在技术和社会协同演进的过程中实现的，而不是单一的技术进步使然（孙启贵等, 2021）。从这个视角看，可信教育人工智能是一种开放、多层次和多要素所构成的新型社会–技术系统。首先，可信教育人工智能是一个多要素组合的复杂系统。技术层面，可信教育人工智能主要由算法、算力和数据组成；社会层面，可信教育人工智能系统的运转需要一系列复杂社会条件的相互配合和支持。其次，可信教育人工智能的创新和应用是一个持续的变革过程。人工智能技术与教育体系协同演进，通过技术将教育体制、教育方法和教育评价等诸多因素有机连接起来，

以实现对教育的重塑。同时,技术的变革路径也会受到这些社会因素的影响。最后,教育的主体是人,这是教育领域与其他社会行业的本质区别,因此,可信教育人工智能需要“以人为中心”。从当前的发展态势看,智能时代下的社会–技术系统的发展更加注重“以人为中心”的设计理念,促使人、机、社会和组织之间更加紧密地交互和协同合作(许为, 2022)。因此,本研究从技术、人和社会三个层面提炼可信教育人工智能的基本框架(如图 3 所示)。

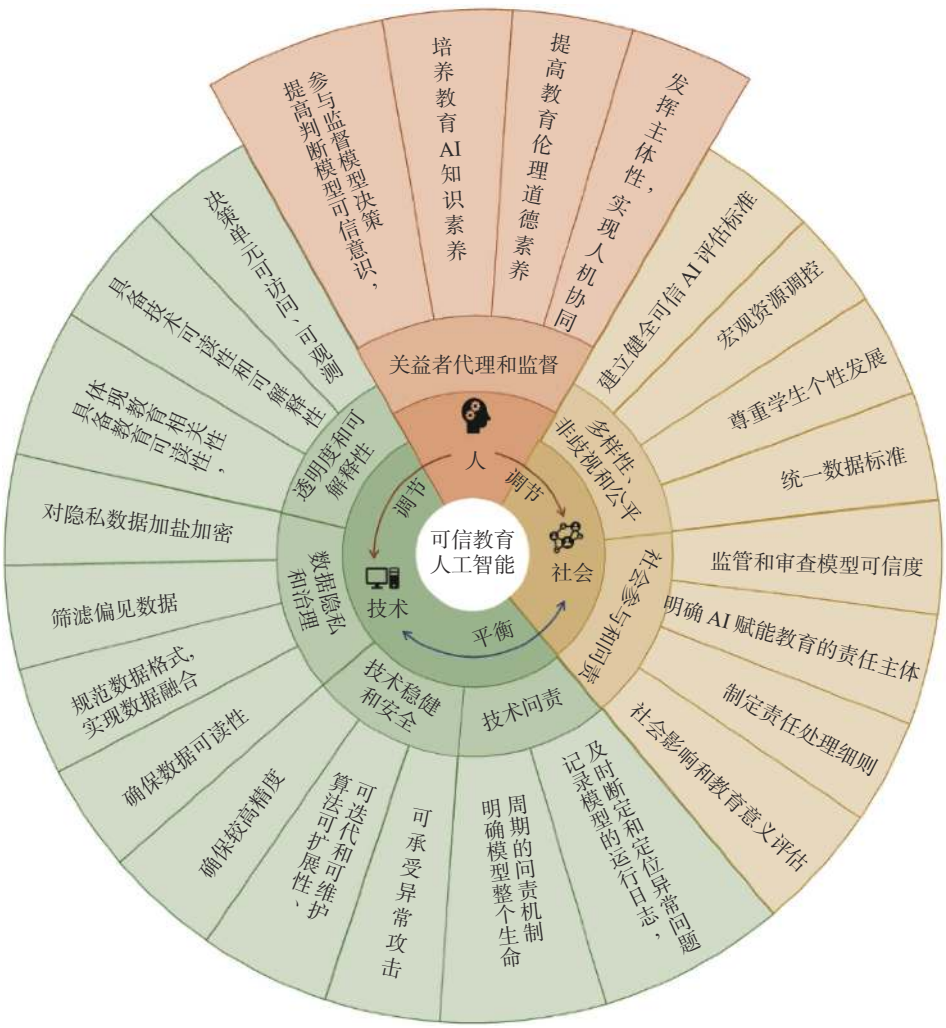

图 3 可信教育人工智能框架

从技术角度来看,可信教育人工智能必须具有良好的透明度和可解释性,能够保障用户数据隐私和进行数据治理,同时必须是稳健的、安全和可问责的。第一,教育人工智能算法模型在整个生命周期都需要实现技术问责机制。在模型投入教育教学场景中使用过程中,要记录从输入到输出的运行轨迹数据,以便于开发人员调试断点,精准发现问题。第二,保证模型的高精度是模型开发的基础要求,在此基础上还需要承受用户并发、数据扰动等情况的攻击,保证完整、可用和保密,同时根据教育相关利益者动态变化的需求,模型能够随时扩展、迭代和维护。第三,教育数据的泄露和窃取从一定程度上会降低用户信任度,因此对个人信息进行特殊处理,除投入模型算法使用外,不作其他用途,且对涉及种族、年龄、地域等可能存在歧视的数据标签进行筛选和过滤,规避数据歧视导致算法推送歧视带来的伦理风险。第四,实现决策单元的可访问、可观测和可追责是保证模型透明度的基本要求,通过可解释机制实现模型的可读性是保证模型可解释性的基本要求。在教育场景下,通过教育视角解释模

型输出的决策，可以保证模型披露的信息能被教育利益相关者理解。

就社会角度而言，需要由各级教育部门和各级学校对教育人工智能的应用进行约束规范，同时提高社会参与度和社会问责意识。在约束规范方面，制定统一的教育数据标准，约束模型开发和使用过程中由于数据偏见带来的算法歧视；在开发和使用人工智能模型过程中始终坚持“以人为中心”的基本原则，尊重学生的个体差异和个性化发展；合理分配不同地域、阶层之间的人工智能技术和资源；建立健全可信人工智能评估标准。在提高社会参与度和社会问责意识方面，相关教育部门可专设部门及时监管和审查教育人工智能的可信度。在教育人工智能模型应用过程中，明确由于非技术人员不当行为造成的教育问题的责任主体，并制定相应的处置办法，同时人工智能技术赋能教育在质而不在量，其功能性、可信度、经济成本都需要进行合理评估。

从人的角度来看，教育主体在教育教学活动中应当充分发挥主体性，不依赖模型的单独作用，而是参与模型决策，共同推进人工智能赋能教学管评的进展。同时，提升教育主体自身的教育公平和伦理道德素养，一方面参与对模型的评估，另一方面对存在问题的决策及时调控。提升教育主体的数字素养，尽管当前可信教育人工智能从技术上难以实现“通俗易懂”，但教育主体仍需要一定的数字素养。人工智能技术在教育场景中的应用非常多样化，自身在选择技术时需要洞察力和辨别力，积极参与并监督模型的决策，避免使用不当而造成负面影响。

四、可信教育人工智能赋能教育数字化转型带来的新机遇

（一）新形态：数据多模态获取、分布式分析、区块链管理

教育数字化转型过程中数字技术的应用以数据为基础，由可信教育人工智能处理数据、解释结果。机器学习的性能与数据量强相关（Halevy et al., 2009），数据量的大小直接影响算法的输出效果，对教育数据的全方位采集不仅可以实现模型的清晰刻画，同时有利于实现模型的可解释性。多模态数据的融合是破解智能教育关键问题的核心驱动力（王一岩, 郑永和, 2022）。多模态数据通常可分为外在行为表现、内在神经生理、人机媒介交互、学习情境活动等（Blikstein & Worsley, 2016; 陈凯泉等, 2019; 牟智佳, 符雅茹, 2021; 王一岩, 郑永和, 2022; Chango et al., 2022），多模态数据的获取是实现多模态数据融合的基本途径，这驱动了记录系统交互日志、追踪脑电心电反应、捕捉眼动及微表情等数字技术参与教育数字化转型。这些数据通过可信教育人工智能的分析、转化、加工和解释，帮助教育参与者更全面深入地剖析学习者的学习过程。

以分布式技术缓解数据压力，以区块链技术强化数据保护。人工智能技术对庞大数据量的需求势必给系统造成负荷，由此驱动分布式存储技术来分担数据压力，进而提高系统的可靠性、易用性和扩展性。智能化数字技术和海量数据资源互相融合推动教育数字化转型，除了可信人工智能外，区块链技术也在一定程度上降低了教育数字化转型过程中可能出现的数据安全、伦理风险问题。区块链技术由多方共同维护，通过使用密码学保证传输和访问安全，在实现数据共享的同时，保护教学利益相关者。结合区块链去中心化的特点，金义富（2017）提出“区块链+教育”来真正实现“以人为中心”的教育需求，张双志和张龙鹏（2020）提出了区块链赋能教育治理结构的技术逻辑，通过区块链中的点对点传输、链式时间戳等技术实现数字信息的共享互通，实现教育主体之间的多元平等、相互信任。

（二）新场景：教育主体从人为干预的“无效参与”过渡到完全可信的“有效参与”

可信教育人工智能优化了数字技术赋能教学管评的业务流程。人工智能治理提出了“人在回路中”的模式，通过对人工智能体赋予特定价值、伦理、道德、意识形态等，让机器“懂人理”（余欣等, 2022），可信人工智能在此基础上进一步实现让人“懂机理”，教育领域中的“回路”体现在人机交互过程中发挥教育利益相关者的主体性。由于人工智能技术赋能教育时存在的“黑箱”问题造成信任问题，在教学环节需要适时人为干预，这种监督式的“无效参与”增加了教育利益相关者的工作成本。可解释性是构建教育领域对人工智能准确认知和良好信心的必然要求（孙波, 2022），可信更是进一步强

化了教育主体对人工智能的信心构建。若能针对教学管评等具体教学流程,构建出被利益相关者接受和信任的可解释方案,就可以实现自动化的教学决策。例如,在个性化教学系统中构建个性化可信方案,能够让学习者明白系统的资源推荐缘由以及取得的何种效果;通过可信的人工智能模型,学习信息管理系统向管理者解释学习者的预测结果,帮助管理者进行教学管理办法的调整和优化;在自动化评测系统中,通过可信人工智能模型以透明、可视化的形式告知评价者具体评价指标,且对评测的具体流程做出解释,从而既保证教育主体适时参与,又减少其劳动负荷,构建人机协同的智能教育结构。当前,国内外已经有不少具备可解释性的学习平台。例如,可解释性 ACSP 系统(Conati et al., 2021)是为学生提供自适应反馈与提示并做出相应解释的自适应导师系统,该系统使用 FUMA 框架进行学习者建模,系统向学习者提供学习反馈的同时,给出判断的理由、预测的原因以及分数计算规则。具备可解释性和可视化的多模态情感分析系统 M2Lens(Wang et al., 2022)的解释引擎采用特征归因方法,构建多层次的模型行为解释,再通过可视化的交互界面呈现给用户,从而实现人机交互。

(三)新生态:数字技术赋能教育数字化转型,数字技术挖掘数据价值,数智融合

可信教育人工智能让教育数据与智能体的融合更加紧密且符合伦理,通过数智融合实现教育数字化转型新生态。教育问题通常由数据反馈,尤其是针对不同的教学场景,教学管评环节关注的数据重点各有不同,比如学习者的学习表现、教学者的教学策略、管理者的治理办法、评价者的评价指标等。可信教育人工智能对这些数据进行融合分析,输出过程性解释和终结性决策,阐明潜在的教育意义。可信教育人工智能要求采集的数据特征粒度更加精细化,来源更加多元化,获取更加精准化,采集方式更加多样化。可信教育人工智能通过缓解责任、公平、安全问题提升教育参与者对其信任度,包括接纳模型决策和反馈真实数据,数据的真实性能够帮助模型得到更加精确的结果,形成良好的人机协同生态,实现数智融合。上海宝山区构建的“未来宝”数字基座使用先进数字技术构建智能教学终端和系统,通过人机交互形成的开放式结构实现了教学资源之间、教学团队之间、各学科之间零距离的沟通与协作,同时通过大量领域知识数据构建的“学科知识图谱”让教学参与者明白资源推送的原因,推动了个性化教学的展开,从而推动基础教育改革,营造数智生态(上海市教育委员会, 2022)。

五、可信教育人工智能促进教育数字化转型的建议

(一)出台可信教育人工智能的标准和法规

将可信视为人工智能设计开发的规范和约束,世界各国出台了一系列的指导建议和管理规范。2019 年欧盟颁布的《可信人工智能伦理指南草案》(Vaggalis, 2019)提出了 10 项构建可信人工智能的要求。2021 年欧盟颁布的《人工智能法案》(European Commission, 2021)提出构建统一规则的可信人工智能监管框架,对高风险的人工智能技术严格监管。同年,中国通信院等(2021)颁布的《可信人工智能白皮书》,聚焦可信人工智能的技术、产业和行业实践,分析其实现路径并提出发展建议。

在教育数据治理领域,我国已出台了相关管理规定,例如《教育部机关及直属事业单位教育数据管理办法》《上海教育数据管理办法》等,明确了教育数据治理的责权及应对数据层面的公平和责任问题,但仍缺少教育人工智能算法设计和使用的管理办法。人工智能赋能教育领域时,既要最大程度发挥人工智能的作用,又要防止技术、数据的滥用和越界。推进可信教育人工智能的立法是从社会层面来解决人工智能的信任问题,通过相关政府部门对可信人工智能提出具体的标准和法规来强化管理,可以缓解教育中的 FATE 问题。例如,在教育数字化转型中建立教育人工智能的算法说明书(曹建峰, 2022),解决透明性问题,提高教育人工智能模型的可信度。通过对算法的详细阐释与说明,判断算法的公平性和伦理性,构建算法问责机制。在教学管评的每一个流程中都有相应的教育主体,同时明确人工智能参与过程中的责任机制,在发现问题之初即可追溯算法出现纰漏的节点,及时干预和制止。

(二)将技术可信度纳入教育数字化建设的评价体系

可信度的量化评估能帮助用户和开发者明确可信程度,有助于实现可信人工智能的健康发展。美

国国家标准与技术研究院发布的《人工智能和用户信任》(Stanton & Jensen, 2021)草案提出了可信度的判断标准，刘晗等人(2022)基于软件的可信属性，从数据、模型和结果三方面构建了人工智能系统可信度量评估框架。教育利益相关者在使用愈发先进但复杂的工具技术过程中，会提出为何要采取这种行动、有什么理论依据、结果是否有效等一系列问题。设计出教育人工智能技术的可信度评估指标，能够帮助教育主体信任、接纳和运用人工智能技术，让教育人工智能融入教育数字化转型中，实现教育流程的高效运作，组成丰富的教育活动形态，形成完备的教育生态。本研究建议，教育人工智能的可信度评估需要在当前可信人工智能的评估框架的基础上结合教育场景的特殊要求，基于前述的可信人工智能原则，从技术、相关利益者、社会三个层面进行量化评估。

（三）深入推进教育数据治理，保护教育数据安全和隐私，构建数据基座

建立有效的数据管理机制是构建安全和有效的数字化教育生态的前提（胡姣等, 2022）。数据标签中的“偏见基因”造成算法决策不公，进而加剧了教育歧视；数字化学习系统没有对采集到的数据进行规范的管理和归档，随着数据量的增大可能会面临数据紊乱；数字化环境中教育利益相关者可能受到恶意攻击和系统干扰或出现人工失误，导致数据泄露；过于注重用户安全的系统可能会建立起数据加密机制(朱嘉文, 顾小清, 2022)，导致各平台之间无法跨越，出现“数据孤岛”的封闭局面。

看似对立的教育数据泄露与数据孤岛问题，实则都是源于对数据的极端处理。在教育数据治理体系的构建上需要“适度”，寻找数据保护与数据融合的平衡点，实现教育数据的智能化集成。对于数据的适度保护，可以通过技术手段和人为干预两种方式。“黑箱”模型的内部结构复杂，未得到合理利用的数据资源反而会造成不必要的泄露。注重数据的产生和应用过程的筛选、过滤，明确数据价值，实现数据保护，辅以可信人工智能的解释来提升技术信任度，从而消弭数据歧视带来的公平问题。同时，相关部门建立健全教育信息的隐私保护制度、提升教育参与者的信息保密意识等也能够形成数据保护机制。对于数据的适度融合，从数据转换和数据规划两方面着手。海量的数字化教育平台承载着形态各异的教育数据，构建可解释的数据挖掘模型对多源异质数据进行提取、剖析和同化，提取特征解释数据关联性，实现跨系统数据的集成。同时，着眼于教育系统本身，在系统设计之初明确数据框架，统一数据格式，帮助系统从数据源头追溯可解释性；也可通过相关的机构或者部门制定统一的数据标准，设定数据管理部门专门处理数据。

（四）提升教育从业者的数字素养，理解和接纳人工智能

数字化转型在一定程度上受制于组织成员的数字化素养(祝智庭, 胡姣, 2022b)。教育从业者，尤其是教师，作为教育数字化转型的重要成员，兼具数字公民和培养数字公民的双重身份(但武刚等, 2022)，必须要提升自身的数字素养，主要有两个途径：理解和信任新技术，接纳和运用新技术。首先，教育从业者对技术的理解有助于构建技术信任。由可信教育人工智能构建的教育教学系统，能够建立起教育者与机器之间的沟通桥梁，更容易被教育者理解，由此打破对机器难以捉摸的偏见，提高掌握数字技术的自我效能感，推动教育从业者进一步接纳和运用新技术。其次，教育从业者接纳并运用数字技术有助于推进教育与技术的融合。从社会–技术系统角度来看，可信人工智能能够提高技术的透明度、增强结果的可靠性、保证运作的安全性，有助于建设更加和平、公正、包容的社会(UNESCO, 2021)。教育从业者是促进技术创新和社会发展的主导力量，教师合理运用可信教育人工智能可以减少劳动成本、提高教学效率、推动技术传播，从而构建一个良性的社会–技术系统。

（江波工作邮箱：bjiang@deit.ecnu.edu.cn）

参考文献

曹建峰. (2022). 人工智能系统可解释性要求的法律规制. *月旦民商法杂志*, (76), 28—39.

陈凯泉, 张春雪, 吴玥玥, 刘璐. (2019). 教育人工智能 (EAI) 中的多模态学习分析、适应性反馈及人机协同. *远程教育杂志*, (05), 24—34.

但武刚, 李玉婷, 王海福. (2022). 高校教师数字素养框架构建与展望. *教育与教学研究*, (09), 41—53.

冯永刚, 赵丹丹. (2022). 人工智能教育的算法风险与善治. *国家教育行政学院学报*, (07), 88—95.
胡姣, 彭红超, 祝智庭. (2022). 教育数字化转型的现实困境与突破路径. *现代远程教育研究*, (05), 72—81.
黄荣怀. (2022). 加快教育数字化转型 推动学校高质量发展. *人民教育*, (Z3), 28—32.
金义富. (2017). 区块链+教育的需求分析与技术框架. *中国电化教育*, (09), 62—68.
荆敏菊. (2015). 中小学生电子产品使用状况及其对心理发展影响与对策的研究综述. *现代教育科学*, (04), 77—79.
刘晗, 李凯旋, 陈仪香. (2022). 人工智能系统可信性度量评估研究综述. *软件学报*, (33), 1—19.
刘三女牙. (2022). 人工智能+教育的融合发展之路. *国家教育行政学院学报*, (10), 7—10.
刘艳红. (2022). 人工智能的可解释性与 AI 的法律责任问题研究. *法制与社会发展*, (01), 78—91.
苗逢春. (2022). 教育人工智能伦理的解析与治理——《人工智能伦理问题建议书》的教育解读. *中国电化教育*, (06), 22—36.
牟智佳, 符雅茹. (2021). 多模态学习分析研究综述. *现代教育技术*, (06), 23—31.
上海市教育委员会. (2022). 上海宝山: 营造数智生态, 推进课堂转型| 基础教育综合改革典型案例. https://new.qq.com/rain/a/20221208A0A9C200
孙波. (2022). 可解释的人工智能: 打开未来智能教育“黑箱”的钥匙. *中国教育信息化*, (04), 3—4.
孙启贵, 汪琛, 王加宇, 叶斌. (2021). 医疗人工智能发展的社会–技术分析与启示. *自然辩证法研究*, (03), 48—53.
托雷•霍尔, 曹梦莹, 明芷安, 袁莉. (2022). 可解释人工智能的教育视角: 基于伦理和素养的思考. *中国教育信息化*, (04), 5—13.
王一岩, 郑永和. (2022). 多模态数据融合: 破解智能教育关键问题的核心驱动力. *现代远程教育研究*, (02), 93—102.
魏亚丽, 张亮. (2022). 从“基于经验”到“数据驱动”: 大数据时代的教学新样态. *当代教育科学*, (02), 50—56.
许为. (2022). 八论以用户为中心的设计: 一个智能社会技术系统新框架及人因工程研究展望. *应用心理学*, (05), 387—401.
杨晓哲, 任友群. (2021). 教育人工智能的下一步——应用场景与推进策略. *中国电化教育*, (01), 89—95.
于聪, 刘飞. (2022). 人工智能教育应用的伦理风险及其对策研究. *机器人产业*, (02), 32—37.
余欣, 朝乐门, 孟刚. (2022). 人在回路型 AI 训练的基本流程与交互模型研究. *情报资料工作*, (05), 34—41.
翟云, 蒋敏娟, 王伟玲. (2021). 中国数字化转型的理论阐释与运行机制. *电子政务*, (06), 67—84.
张坤颖, 张家年. (2017). 人工智能教育应用与研究中的新区、误区、盲区与禁区. *远程教育杂志*, (05), 54—63.
张双志, 张龙鹏. (2020). 教育治理结构创新: 区块链赋能视角. *中国电化教育*, (07), 64—72.
中国通信院, 京东探索研究院. (2021). 可信人工智能白皮书. http://www.caict.ac.cn/kxyj/qwfb/bps/202107/t20210708_380126.htm
朱嘉文, 顾小清. (2022). 打通“数据孤岛”实现数据互联互通. *教育传播与技术*, (04), 3—8.
祝智庭, 胡姣. (2022a). 教育数字化转型的实践逻辑与发展机遇. *电化教育研究*, (01), 5—15.
祝智庭, 胡姣. (2022b). 教育数字化转型: 面向未来的教育“转基因”工程. *开放教育研究*, (05), 12—19.
祝智庭, 彭红超, 雷云鹤. (2018). 智能教育: 智慧教育的实践路径. *开放教育研究*, (04), 13—24+42.
Blikstein, P., & Worsley, M. B. (2016). Multimodal Learning Analytics and Education Data Mining: Using computational technologies to measure complex learning tasks. *Journal of learning Analytics*, 3, 220—238.
Brence, F. , & Mauhart, J. (2019). Digital Enablement: Turning Your Transformation Journey into a Successful Journey. https://www2.deloitte.com/content/dam/Deloitte/at/Documents/human-capital/at-digital-enablement-turning-your-transformation-into-a-successful-journey.pdf
Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *WIREs Data Mining and Knowledge Discovery*, 12(4), e1458.
Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298, 103503.
Duggan, T., & Corporation, T. (2020). AI in Education: Change at the Speed of Learning. https://iite.unesco.org/wp-content/uploads/2021/05/Steven_Duggan_AI-in-Education_2020-2.pdf.
European Commission, & Directorate-General for Communications Networks, C. and T. (2019). Ethics guidelines for trustworthy AI. *Publications Office*.
European Commission. (2021). Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research policy*, 31(8-9), 1257—1274.
Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2), 8—12.
Hanna, N. (2018). A role for the state in the digital age. *Journal of Innovation and Entrepreneurship*, 7(1), 5.
IEEE Computer Society. (2022). 2022 Technology Predictions. https://ieeecs-media.computer.org/media/tech-news/tech-predictions-report-2022.pdf

Kizilcec, R. F., & Lee, H.（2020）. Algorithmic Fairness in Education. *CoRR*, abs/2007.05443.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., & Tang, J.（2022）. Trustworthy AI: A Computational Perspective. *ACM Trans. Intell. Syst. Technol.*

Stanton, B. , & Jensen, T.（2021）. Trust and Artificial Intelligence. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf

Trump.（2020）. Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/

UNESCO.（2021）. 人工智能伦理问题建议书. https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi

UNESCO.（2022）. United Nations Transforming Education Summit——Thematic Action Track 4 on ‘Digital learning and transformation’. https://transformingeducationsummit.sdg4education2030.org/track/digital

Vaggalis, N.（2019）. Ethics Guidelines For Trustworthy AI. https://www.i-programmer.info/programming/artificial-intelligence/12702-ethics-guidelin

Wang, X., He, J., Jin, Z., Yang, M., Wang, Y., & Qu, H.（2022）. M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 802—812.

注　释：

①https://www.ibm.com/topics/digital-transformation#anchor-83353465

②https://www.microsoft.com/en-us/research/theme/fate/

（责任编辑　孙世杰）

The Core Technology Engine of Digital Transformation in Education: Trustworthy Education Artificial Intelligence

Jiang Bo　Ding Yingwen　Wei Yuang

(Department of Education Information Technology /Institute of AI Education, SH, East China Normal University, Shanghai, 200062, China)

Abstract: The digital transformation of education aims to reengineer the education and teaching process and improve the quality and efficiency through digital technology. On the one hand, digital technology represented by artificial intelligence as the technical engine of digital transformation of education drives the continuous deepening of digital transformation of education. On the other hand, the “black box” problem of artificial intelligence technology has caused a crisis of human-machine trust, and there is a risk of violating basic constraints such as fairness, accountability, transparency, and ethics in education, hindering the digital transformation of education. This study sorts out the four major governance problems caused and aggravated by artificial intelligence to help the digital transformation of education, analyzes the theoretical research and development status of trustworthy artificial intelligence in education, puts forward the basic framework of trustworthy artificial intelligence in education, summarizes the opportunities brought by trustworthy artificial intelligence in education for the digital transformation of education, and puts forward the development suggestions of trustworthy education artificial intelligence to promote the digital transformation of education. The research recommends the introduction of relevant standards and regulations for trustworthy educational artificial intelligence, incorporating technical credibility into the evaluation system of education digitalization, further promoting educational data governance, and improving the digital literacy of education practitioners.

Keywords: digital transformation of education; trustworthy artificial intelligence; trustworthy artificial intelligence in education; governance of education