可信赖人工智能教育应用的建设路径与现实启示
——以英国典型举措为例

□沈苑　胡梦圆　范逸洲　汪琼

摘要：人工智能技术在辅助教学上展现出了巨大潜力，也因其不透明性、无法预测性以及不当使用而引发争议，建设可信赖的人工智能教育应用（AIED）成为国际共识。可信赖AIED强调“以人为本”的价值宗旨，以可靠、安全的AI技术为底层基础，以合伦理性的理解和使用为关键保障，观照教育情境的复杂性，摒弃将通用AI伦理原则或其他AI应用领域的专业伦理原则生搬硬套到教育领域。英国作为人工智能伦理治理的先行者，在可信赖AIED建设方面积累了丰富经验，尤其体现在《人工智能保障生态系统路线图》中。该《路线图》从技术的开发、采购与使用三大阶段出发，分别采取伦理设计、伦理核查和AI素养培养等具体策略，构建了可信赖AIED的保障体系。借鉴英国AIED的典型举措，我国应在开发阶段对智能教育产品开展本土化的价值敏感设计，采购阶段对智能教育产品的可信赖度做出有意义的解释，使用阶段对师生用户的人工智能素养加以培养和提升，以推动技术开发者、产品采购者、一线使用者等利益相关方共促可信赖AIED的本地建设。

关键词：人工智能教育应用；可信赖人工智能；伦理设计；英国实践

中图分类号：G434　**文献标识码：**A　**文章编号：**1009-5195(2023)04-0065-10 doi10.3969/j.issn.1009-5195.2023.04.008

基金项目：教育部科技司2022年教育领域智能社会研究“智能教学环境人机合作平衡点探查”（D2022010）；2022年江苏高等教育学会《江苏高教》专项课题“人工智能高等教育应用伦理研究”（2022JSGJKT001）。

作者简介：沈苑，博士研究生，北京大学教育学院（北京　100871）；胡梦圆，硕士研究生，北京大学教育学院（北京　100871）；范逸洲，博士，助理教授，硕士生导师，北京大学教育学院（北京　100871）；汪琼（通讯作者），博士，教授，博士生导师，北京大学教育学院（北京　100871）。

一、引言

近年来，人工智能教育应用（Artificial Intelligence in Education，AIED）为教育形态的重塑提供了巨大的可能性。尤其是以ChatGPT为代表的新一代人工智能，更是凭借其高生成性、高自主性、高交互性，对传统的教育方式和教学方法带来冲击，并进一步推动“因材施教”“终身学习”等教育理想走向现实。同时，因机器学习具有“黑箱”特征，与之相关的安全性和透明度等伦理质疑也开始蔓延（Dignum，2019）。总体来看，AIED对于传统课堂来说尚属陌生的“外来物”，在师生信赖不足的情况下，不乏技术误用、滥用、弃用等情况出现。为保障AIED充分发挥其功能优势并为教育赋能，亟须建设起符合伦理要求且值得师生信赖的智能生态系统。基于此，探究可信赖的人工智能教育应用（Trustworthy AIED）成为教育领域重要的研究话题。

长期以来，英国一直都是人工智能伦理治理的先锋。作为“人工智能之父”艾伦·图灵（Alan Turing）的故乡和诸多著名AI企业（如DeepMind）的发源地，英国对AI伦理问题在政策上给予了大量关注。例如，在《人工智能行动计划》（AI Action Plan）中，英国政府将“有效治理”作为国家发展AI的三大目标之一。2021年，英国中央数字与数据办公室、人工智能办公室与内阁办公室联合发布了《自动决策系统的伦理、透明度与责任框架》（Ethics，Transparency and Accountability Framework for Automated Decision-Making），提出了保障系统安全、可持续、伦理性的7项要素（GOV.UK，2021）。同年出台的《人工智能保障生态系统路线图》（The Roadmap to an Effective AI Assurance Ecosystem，下文简称《路线图》）明确提出了各利益相关方在建设可信赖AI过程中应分别承担的责任

(CDEI，2021a)。近年来，英国的学术组织、研究机构、学校也都在努力探索可信赖AIED的建设路径，贯彻相关政策的核心精神，积累了丰富的实践经验。有鉴于此，本文将首先对可信赖AIED的发展和建设背景进行分析，然后针对英国在AIED全生命周期中开展伦理治理的具体实践与典型案例进行剖析，最后从AIED设计者、采购者和一线用户等利益相关方的视角出发，反思和总结英国相关举措对于我国可信赖AIED本土建设的启示。

二、可信赖AIED的发展背景

“可信赖AIED”由“可信赖AI”这一概念引申而来，梳理其内涵发展、现实困境和领域特征，有助于人们达成对可信赖AIED的一致共识。

1.可信赖AI的内涵发展

我国学者李应潭在《人工智能可信赖性与可信赖算法研究》一文中将“可信赖AI”描述为“在没有硬件故障的情况下，对于向它提出的一批问题，能够做到没有把握或无力回答的就不回答，有把握的才回答，回答的则保证正确”（李应潭，1999）。21世纪以来，这一概念强调的重心逐渐从技术维度转向伦理维度，成为了社会技术伦理治理的最终目标之一。国际标准组织/国际电工委员会的第一联合技术委员会（ISO/IEC JTC1）将人工智能的“可信赖”界定为“有能力提供客观证据证明AI产品或系统可以完成特定要求以满足利益相关者的期望”（ISO/IEC TR 24028：2020，2020）。

英国政府在人工智能相关战略制定方面始终非常强调人工智能的“可信赖”问题。2016年，英国政府发布《人工智能：未来决策制定的机遇和影响》（Artificial Intelligence：Opportunities and Implications for the Future of Decision Making）报告，指出“公众信赖”是人工智能得到有效利用的必要条件，建设信赖的核心在于公开的对话（GOV.UK，2016）。2017年，《在英国发展人工智能》（Growing the Artificial Intelligence Industry in the UK）报告再次指出，建立公众对AI的信赖对于英国人工智能的成功发展至关重要，特别要增强数据信赖（Data Trust），以促进数据持有方和技术开发方之间安全地共享数据；该报告还要求行业与政府携手开发经过验证的框架和协议以确保数据共享的安全性和互惠性（Hall et al.，2017）。2018年4月，英国上议院下属的人工智能委员会发布《英国人工智能发展的计划、能力与志向》（AI in the UK：Ready，Willing and Able），指出如果公众被过度暴露于负面的、失真的AI描述中，可能会导致公众反弹，同时承认某些人工智能应用本身可能涉及误导和欺骗，应辩证看待AI技术（House of Lords，2018）。此外，英国数字、文化、媒体和体育部（DCMS）在《国家人工智能战略》（National AI Strategy）中进一步重申要建设“世界上最受信赖和支持创新的人工智能治理系统”的愿景（Kwarteng et al.，2021）。

2019年4月，欧盟人工智能高级专家组发布的《可信赖人工智能伦理准则》（Ethics Guidelines for Trustworthy AI），可谓是可信赖AI发展道路上的一个里程碑，其确立了“可信赖AI”的三项必要条件，即人工智能须符合法律法规、人工智能须满足伦理道德原则及价值、人工智能在技术和社会层面应具有可靠性。该准则还明确了构成可信赖AI的7个关键要素，分别为人的能动性与监督，技术稳健性与安全性，隐私与数据管理，社会与环境福祉，多样性、非歧视性与公平性，透明性和问责制度（AI HLEG，2018）。聚焦于教育领域，经济合作与发展组织（OECD）于2020年4月发布了《教育中的可信赖人工智能：前景与挑战》（Trustworthy Artificial Intelligence （AI）in Education：Promises and Challenges），提出了可信赖AI在教育中的要求，即只有同时满足“AI能准确地执行任务”和“以公平且恰当的方式被使用”这两项条件时才能取得公众信赖（Vincent-Lancrin et al.，2020）。

综上，可信赖AI的内涵经历了从物理层面的准确性延伸至社会层面的合伦理性的发展过程。开发出可靠的AI技术，确保技术稳健安全和透明，始终是信赖建设的基础。而利益相关者如何能够辩证地理解和恰当地使用AI，包括正确地管理数据、包容地推广技术、主动地监督使用，则是巩固信赖的保障。

2.AIED面临的信赖危机

目前，AIED产品已被广泛应用于教育教学中。沉浸式体验系统、智能导师系统、自动评估系统的出现，为传统的教学模式注入了变革的动力。已有很多原本不是专门为教育学科设计的各种社交网络、博客、游戏平台和移动应用程序，也都在潜移默化地影响着学习生态系统。但从应用现状来看，机器学习所带来的无法预测性、不透明性，以及对

人工智能的不当使用等因素，导致教育环境中频繁出现人机信赖危机。

比如，2020年受新冠疫情影响英国高考无法正常开展，英国资历及考试评审局采用了智能系统为考生打分，但经统计发现，约39.1%的考生的成绩被系统低估，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（Rahim，2020）。2022年秋季美国高校有学生借助OpenAI开发的聊天机器人ChatGPT做作业、写论文，甚至得到了高分，更是引发了高校的警觉，使一线教师陷入深深的担忧之中（Roose，2023）。国内方面，有学校使用头环监测学生的课堂注意力，也有学校使用手环记录学生的位置和行为，并将这些数据报告给家长和教师，致使“校园监狱化”等负面论断频频出现（新京报，2019）。

“以人为本”是教育实践者与研究者开展全部工作的起点与终点。上述事件反映出，目前应用于教育领域的人工智能产品，如果忽视了教育场景下的伦理诉求，非但不能达成预期效果，反而会致使教育异化，引发公众抵触情绪和人机信赖危机。美国教育传播与技术协会（AECT）在2008年明确将教育技术界定为“通过创建、使用和管理适当的技术和资源来推动学习和提升表现的研究与符合伦理的实践”（Januszewski et al.，2013）。上述定义中的合伦理性意味着对AIED的伦理审视刻不容缓。

3.教育场景的复杂性

AIED如何设计和使用也受到教育场景复杂因素的影响。首先，教育场景下的受教育对象往往是易受外界影响的儿童和年轻人，其身心尚未成熟，相较于成人用户更需要正当的支持引导。其次，与其他AI应用领域不同，教育领域中的很多问题并非“生死一线”问题，一项不恰当的教育决策看似不如医疗事故或交通事故那么严重，但这也意味着利益相关者很难及时发现决策的问题所在，因此也需要付出更多成本来修正或弥补其后果。再次，相较于制造、零售、安防等AI应用较为成熟的行业，人工智能在教育领域的应用仍处于摸索阶段，其带来的直接和间接影响仍缺乏有效验证，而且因为教育实践本身的复杂性和隐晦性，以及影响教育目标实现的因素众多，致使很难轻易地将技术因素剥离出来分析其影响路径。最后，教育环境中的行为主体有自己的价值倾向性，难以一概而论，就像多数人工智能学家所追求的系统准确度并不一定是利益相关者最看重的东西。比如，学生可能希望在学习上拥有更多自主性，教师则希望在学习过程中保持人工监督和随时干预，而家长则希望系统决策是透明可解释的（Holmes et al.，2021）。

基于教育领域的特殊性与复杂性，建设可信赖AIED并不能简单地将通用的AI伦理原则或其他AI应用领域的专业伦理原则“生搬”到教育领域，而是要兼顾技术和教育的双向视角，挖掘突破AIED场景中伦理困境的内在进路。

三、可信赖AIED的建设路线

随着AI的快速发展及其教育应用场景的不断拓展，将碎片化的治理规则和治理工具整合为保障生态系统的迫切性和必要性持续凸显。专门为英国政府提供人工智能安全、伦理和创新发展等政策建议的智库——英国数据伦理与创新中心（CDEI）于2021年12月8日发布《人工智能保障生态系统路线图》（下文简称《路线图》），明确提出“在未来五年内要建设一个蓬勃发展和有效的人工智能保障生态系统”的目标，并提出了较为成熟的建设路线（CDEI，2021a）。下文概述了《路线图》中对于AI要达到“可信赖”标准所提出的具体要求以及对应的行动指南，可作为教育领域建设“可信赖AIED”的重要参照。

1.建设可信赖AIED的双重要求

《路线图》将建设可信赖AI遇到的挑战划分为信息问题和沟通问题两类，前者指向“人工智能系统的设计是否值得信赖”，后者指向“人工智能系统的使用是否值得信赖”。

信息层面的信赖建设聚焦于提升技术产品自身的稳健和可靠程度。如果AIED系统存在数据集质量欠佳、算法偏见、决策机制不透明、隐私保护不当、无法抵抗外部攻击等来自技术局限或设计不周的表现，则会削弱公众信赖，进而引发其对智能产品的抵触和质疑。有研究者发现，IBM、旷视、微软的人脸识别产品识别男性的准确率均高于女性，这类存在偏见的算法模型如果被用于教育中，很可能会加剧男童与女童在技术可及性方面的鸿沟（Buolamwini et al.，2018）。

沟通层面的信赖建设聚焦于提升技术在部署和使用过程中的恰当程度。这取决于用户对技术的信赖与技术的可靠性是否匹配，过多和过少的信赖都会阻碍教育目标的实现。设想如果某著名的技术企

业推出一个智能教学系统，即便此系统的算法并不成熟、准确度也无从考量，但师生可能会因为公司名气大或宣传力度大而盲目相信此系统，这种信赖错置可能会导致误用和滥用。而如果有一个算法成熟、准确度高的系统，但由于其出自一家不知名的初创公司，或技术开发者没有准确地将证明它很可靠的证据传达给师生，或学校领导和师生缺乏对产品背后AI技术原理的了解等，都可能导致用户不愿意信赖和使用该系统，从而导致优质资源被浪费。

2.建设可信赖AIED的多方行动

《路线图》还列明了处于人工智能供应链不同位置上的关键利益相关方，包括AIED开发人员、采购人员、一线用户以及受AI影响的其他个人，尤其针对前三方还分别提出了明确的行动指南（见图1），即要求多方在人工智能全生命周期的各个阶段协作配合，以建设起人工智能保障生态系统。

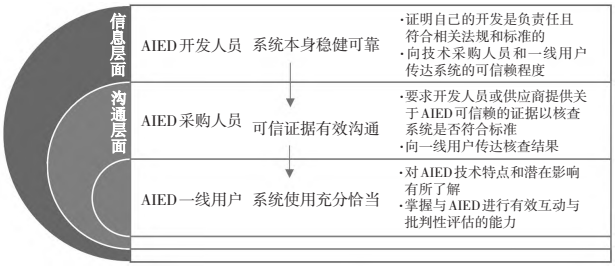

图1　可信赖AIED保障生态系统的三方行动指南

（改编自CDEI，2021a）

具体来说，AIED开发人员需要保障技术本身的稳健可靠，遵循相关法规标准进行开发，尽可能地规避数据泄露、偏见歧视、损害自主等技术伦理问题。AIED采购人员需要对技术的可信程度与希望通过此系统达成的教育目标的匹配程度进行有效评估，并承担起促进多方沟通的任务，避免信息不对称情况的出现。AIED一线用户需要提升自身的技术素养，能根据技术的可信程度和相关证据在实践中灵活、恰当地使用AIED。

从目前英国在教育领域的具体实践来看，AIED的开发人员、采购人员和一线用户都在推动可信赖AIED的建设上进行了初步探索。后文将分别从开发保障、采购保障与使用保障三个角度进一步解读英国在建设可信赖AIED的典型举措，并就其对我国教育领域人工智能伦理治理的启示展开讨论。

四、开发保障：AIED产品的伦理设计

按照《路线图》的要求，AIED的开发者需要首先证明自己的开发是负责任且符合相关法规和标准的，同时要不断向技术采购人员与一线用户传达其系统的可信赖程度。

在2021年英国政府网站发布的《AI晴雨表第五部分——教育》（AI Barometer Part 5 – Education）中，将AIED划分为面向学生、面向教师和面向区域三类（CDEI，2021b）。在这些不同类别的AIED产品开发中，设计者都非常重视提升技术的可信赖程度。下文选取了英国3个典型的AIED开发项目，根据《可信赖人工智能伦理准则》的7项要素对其伦理重点进行归类，并对其设计策略进行详细解读（见表1）。

表1　英国AIED产品的伦理设计策略

AIED类型	产品/项目名称	伦理重点	设计策略
面向学生	AIAccess	透明性	部署开放学习者模型
面向教师	ClassCharts	隐私与数据管理	遵循数据保护规范
面向区域	BIT学校审查项目	多样性、非歧视性与公平性	建设高质量的数据集

1.面向学生的AIED：以AIAccess为例

面向学生的AIED往往以“智能导师系统”或“适应性学习平台”的形式出现（Baker et al.，2019）。其通常以增进学生个性化学习为开发目标，具备基于学生需求设计和推送学习材料、诊断学生知识中的强项与薄弱处、提供自动化反馈、促进学习者合作等功能。

透明性是此类产品开发中关注的关键要素，其强调“算法的可解释性”，即要求AI系统的工作原理、工作机制是可以被解释的，系统所做出的任何决策都是有迹可循的。一个不透明的AI“黑箱”会使教师与学生难以深入理解并接受系统所进行的评估、推荐或预测，因而可能导致教师做出不当决策，学生错失进一步提升学习的机会。

由伦敦大学学院知识实验室（UCL Knowledge Lab）开发的人工智能评估系统软件AIAccess，通过使用开放学习者模型（Open Learner Model，OLM）以保障其透明性。AIAccess是一款帮助学生进行数学与科学学习的智能评估软件，能以可视化地图的方式展示系统对于学习者完成任务表现的判断，包括学生正在学习的内容在整体学科知识地图中的位置、学生在每个话题下已经完成的不同难度水平的学习内容，以及所获得的系统支持程度等（Luckin，2017）。OLM使得系统的推理过程更加透明：一方面学生能够更多享有对自己学习的控制权，增强他们对自身学习负责的意识（万海鹏等，2021）；另一方面

也让教师能够更好地掌握学生的知识理解水平，以便进一步开展个性化的学生管理与教学。

2.面向教师的AIED：以ClassCharts为例

面向教师的AIED旨在帮助教师减轻管理与教学方面的压力。在这类系统中，通常会通过自动发布学习任务与材料以减轻教师的工作负担；或者通过对学习者行为进行预测分析，以辅助教师更好地进行针对性的分组教学与班级管理。

隐私和数据管理是此类产品开发中关注的关键要素。教师在进行决策的过程中，不可避免地要收集学习者多方面的数据，这就要求保障技术足够稳健安全，以避免侵犯学生个人隐私、泄漏个人数据等风险。

ClassCharts是一款辅助教师进行课堂管理的人工智能工具，能够监测教室中学生的行为，基于学生间的互动历史与相互影响即时生成学生座位排布，追踪教室中特定学生间的互动，并快速生成学生行为分析报告，以便于教师高效管理学生的行为。这些功能的实现需要收集众多学生的行为与互动数据。为确保用户安心使用，ClassCharts所属公司Edukey声明会严格遵守欧洲《通用数据保护条例》（General Data Protection Regulation，GDPR），以标准化的政策和流程管理用户数据。在数据隐私方面，公司声明学校是教师与学生相关数据的控制方，具有决定收集何类数据、由谁处理数据的权力，而公司只是数据的处理方而非控制方，并且公司会保障用户的“被遗忘权”（The Right to Be Forgotten），即用户删除个人信息的权力。在数据安全方面，公司指出他们用于存储和处理数据的谷歌云服务器位于英国境内，从而保证了用户数据存储在欧洲经济区范围内，并会在平台中通过加密技术、防火墙等多层保护措施保障用户数据的安全（Edukey，2022）。

3.面向区域的AIED：以BIT学校审查项目为例

面向区域的AIED多以学校管理系统的形式出现，主要用于学校与地区层面上的整体改进，涉及不同学校间的数据共享。

多样性、非歧视性与公平性是此类产品开发中关注的关键要素。在数据共享的过程中会涉及多元化的机构和人群，内嵌于技术的偏见问题可能会加剧对特定种族或社会文化背景人群的不公正对待，加深学生在资源和机会获取方面的差距，进而扩大阻碍教育公平发展的数字鸿沟。

英国政府成立的行为研究小组（Behavioural Insights Team，BIT）近年来开展了借助机器学习为学校表现打分的大型社会实验，以帮助英国教育标准局等机构进行科学决策。该评估算法采集了大量的教育数据，如各学校本地数据库中的学生数据、英国教育标准局记录的学校数据、学校职工普查数据、家长调查数据等，并基于这些数据对学校的未来表现进行预测（Baker et al.，2019）。但是，这种做法可能会因为数据样本偏差而致使算法模型偏见，比如拥有黑人学生或教师比例更高的学校可能更容易被判断为具有高风险的学校。对此，项目负责人声明BIT团队在项目规划初期就将偏见风险纳入考量，在创建数据集时特别加强了对数据集的来源、数据规模、代表性、特征选取和标注质量的把关，尽可能选取更高质量的数据集；而对于可能直接或间接引发宗教、种族偏见等的数据则进行多轮审查与清洗，以期尽可能保障对机构及其师生的公平对待（Reynolds，2017）。

五、采购保障：AIED产品的伦理核查

按照《路线图》的要求，教育机构负责采购AIED的人员应当要求开发人员和供应商提供关于技术值得信赖的证据，以确保他们采购的系统符合标准，并向一线用户传达自己评估的结果。英国人工智能教育伦理研究所（The Institute for Ethical AI in Education，IEAIED）提出了面向AIED采购者的伦理框架（下文简称IEAIED框架），要求采购者在技术采购和部署阶段保持伦理敏感度，担负起教育机构与技术企业的沟通桥梁责任（IEAIED，2021）。

1.IEAIED框架的伦理维度

IEAIED框架包含9个伦理维度和33条对应的核查指标，实现了对AIED开发目标以及目标实现形式的全方位审查。9个伦理维度如下：

一是实现教育目标，即AIED应被用于实现具有可靠证据支持的、以学习者为中心的明确的教育目标。二是评价形式，即AIED应被用于评价和识别更广范围内的学习者能力。三是管理和工作量，即AIED应能增强教育机构的能力，同时尊重人类关系。四是平等，即AIED应被用于促进不同学习者群体之间的公平，而不是以歧视任何学习者群体的方式被使用。五是自治，即AIED应被用于提高学习者对其学习和发展的控制水平上。六是隐私，即AIED应在隐私和合法使用数据来实现教育目标之间取得平衡。七是透明和问责，即人类最终要对

教育成果负责，因此应对AIED的运作方式进行适当的监督。八是知情参与，即学习者、教育者和其他相关从业者应对AIED及其影响有合理的理解。九是合乎伦理的设计，即AIED应由了解AIED潜在影响的人来设计。

上述9个维度要求教育机构的采购者不仅要对AIED产品本身进行审查，也强调了机构内部要进行自我反思和审视，比如前三个维度要求学校本身要知道自己使用AI的目的究竟是什么，以及明确AI能为学校带来的益处和风险。此外，IEAIED框架也要求采购者与开发AIED的研发团队进行深度沟通，比如最后一个维度要求采购者了解开发AIED的团队构成以及成员的技术素养水平。

2.IEAIED框架的独特优势

与其他现有的AIED伦理框架相比（Aiken et al.，2000；Holmes et al.，2021），IEAIED伦理框架在伦理视角、受众群体、可操作性方面都具有鲜明的优势。

第一，在伦理视角方面，已有框架大多是从人工智能伦理原则视角出发将权威的技术原则（如计算机协会道德与职业行为准则、阿西莫夫机器人三定律、罗杰克拉克机器人七原则、欧盟可信赖人工智能伦理准则等）演绎于教育领域，而IEAIED框架中不仅包含了平等、自治、隐私、透明和问责、知情参与、合乎伦理的设计等技术伦理要求，还更多地关注到了教育目标、评价形式、管理和工作量等教育领域的特定伦理要求。

第二，在受众群体方面，已有框架大多没有区分特定的面向对象，涉及内容既包括技术本身要满足的标准，也包括使用过程中人要注意的事项。而IEAIED框架直接面向教育机构中负责采购AIED产品的负责人，也间接地要求技术开发者提供足够的信息给采购者。

第三，已有框架大多包含了AIED需要达到的普遍标准，但没有作出阶段性的要求，也缺少操作指南，而IEAIED框架还为采购者提供了极为详细的核查清单，包含了产品采购前、采购时、部署后每个阶段应该核查的内容，具有更强的针对性，也更能保障该框架的落地实施。

六、使用保障：提升一线用户的AI素养

根据《路线图》要求，作为一线用户的师生需要对智能系统有所了解，以便恰当地运用AIED来达成教育目标。近年来英国对教育的关注点放在了提升师生AI素养上。AI素养包括了解AI的基本技术和工作原理、批判性地理解AI对现实生活的影响和潜在的伦理问题、进行AI产品的创造等方面。《路线图》指出，让每名学生都掌握基本的AI知识，不仅是理解基础的编程、推理或数学概念，也不仅是简单的伦理原则，而是要让学生成为“有意识的、有自信的AI使用者”，知道应当问什么问题、应当注意哪些社会风险和伦理问题，以及AI能够带来什么机会。同时，《路线图》也支持每位教师与AI同进步，帮助他们掌握有关AI技术的基本信息，以便在教学实践中做出最优选择。

1.英国基础教育阶段的AI素养提升

在基础教育阶段，英国政府与社会组织协力提升师生的AI素养。在学生素养方面，英国的独立慈善组织Apps for Good为英国超过20万名中小学生提供了免费的技术课程，在帮助学生广泛了解技术运用原理，并指导学生基于现实问题自主进行智能产品开发的同时，还讲授了关于AI算法使用的社会、法律与伦理影响等内容，以增进学生对可信赖AI的了解，旨在让他们成为具有批判能力和负责任的AI技术使用者。在教师素养方面，由英国教育部资助建立的英国国家计算机教育中心（National Centre for Computing Education，NCCE）为英国中小学教师提供了具有学分与证书认证的计算机与技术课程以及免费的教学资源，并为有需求的学校提供资金支持，其目标是通过政府对NCCE项目的支持，让英国每所学校的学生都享有世界领先的计算机教育。

2.英国高等教育阶段的AI素养提升

在高等教育阶段，英国大力发展数据科学与人工智能学科，以及人工智能伦理教学项目。目前已有92所英国大学提供了关于人工智能的本科生课程（UCAS，2022），并且有超过224个有关人工智能的研究生项目（FindAMasters，2022）。在英国政府发布的《在英国发展人工智能产业》（Growing the Artificial Intelligence Industry in UK）报告中还提出，英国大学应当结合雇主与学生的需求，为更多本科不是计算机或数据科学的学生开设更多人工智能一年制转修硕士课程（One-Year Conversion Masters Degrees）。该报告还提出在一流大学还需创造至少200个AI领域的博士学位名额，而且预计在2025年需要增加至少1000个人工智能博士学位名额（Hall et al.，2017）。

此外，英国的多所大学还开设了关于AI伦理的课程。如剑桥大学提供了“AI伦理与社会”

(AI Ethics and Society)硕士项目，旨在培养学生在AI技术所带来的机遇与挑战下所必备的专业知识与技能，从而让他们能够理解与应对复杂社会技术系统中关于隐私、监控、公平、算法偏见、负责任的创新等方面的问题(Leverhulme Centre for the Future of Intelligence，2022)。牛津大学也成立了AI伦理研究院，重点关注AI与民主、人权、环境、治理、人类幸福、社会6个议题(Institute for Ethics in AI，2022)，以促进可信赖AI的设计与使用。

七、英国可信赖AIED建设的经验与启示

英国在探索可信赖AIED道路上积累了丰富的经验，《路线图》中的三方行动指南以及前文述及的典型做法为我国可信赖AIED建设提供了重要的借鉴。考虑到中西方在文化价值观以及AIED普及程度上的差异，我国也应该结合本土实际情况，灵活地调整可信赖AIED的建设路线和具体举措。

1.开发阶段：开展AIED本土化价值敏感设计

《路线图》中始终强调多方利益相关者的合力建设，要求在技术设计阶段就要确保AIED有足够的稳健性。这种倡议与芭提雅·弗里德曼(Batya Friedman)等人所提出的价值敏感设计理论(Value Sensitive Design，VSD)相契合。在价值敏感设计的视角下，需要将利益相关者的伦理价值嵌入AIED的设计之中，通过伦理要素制约设计，达到使技术服务于正向度的社会建构的目标(Friedman et al.，2002)。如前文中采用开放学习者模型、遵循隐私保护规范、采用高质量数据集等做法都是价值敏感设计的实现形式。

以“隐私”这项伦理价值为例。AIED的开发者在进行产品设计时就要认识到，在便捷地收集数据的同时自身也承担了更大的数据保护责任，应当依托国家政策制定严格的数据使用规范。欧洲GDPR就为众多信息技术产品划定了不可逾越的隐私边界，如ClassCharts这款软件的开发商就明确表示完全遵循GDPR的要求。对于我国来说，AIED在隐私方面的价值敏感设计则需要对标近年来我国出台的《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等政策要求，对数据收集范围、收集过程、数据处理过程、数据所有者、可获取数据者、知情同意流程等方面加以清晰界定和严格执行。

另一方面，我国开展AIED价值敏感设计也需要契合我国本土文化价值的特点。比如，IEAIED要求“在隐私和合法使用数据以实现教育目标之间取得平衡”。这种平衡点的位置在不同文化背景下很可能会有所差异。参考霍尔的个人空间理论，有研究者指出，我国的传统文化更强调集体主义，更能包容拥挤的个人空间，更愿意牺牲一部分隐私换取交际关系或者其他效用(郭望舒，2009)。以校园人脸识别技术为例，2019年瑞典的一所中学因为使用这类系统进行考勤而被瑞典数据保护局认定为违反GDPR，并处以2万欧元的罚款(Coraggio，2019)。美国多个州市因为校园暴力事件频发，近年来才开始允许家长申请安装教室摄像头(Heintzelman et al.，2017)。相较于一些西方国家极为严苛的隐私保护手段，我国教育领域对于这类技术则采取了更加灵活、开放的态度，目前已有许多中小学和高校借助人脸识别技术开展智慧校园建设，在承认此类技术优越性的同时，通过《教育部等八部门关于引导规范教育移动互联网应用有序健康发展的意见》等政策法规加以制衡。因此，在产品设计过程需要增强对我国本土文化的审视，切忌完全效仿国外做法而导致“水土不服”。

2.采购阶段：沟通过程中对产品的可信赖度作出有意义的解释

英国IEAIED所提出的伦理框架启示我们要关注AIED的采购阶段。我国教育机构负责采购AIED的人员应该从多个维度对产品进行审查，要求AIED的设计者提供关于产品可信度的一系列证据，以说明AIED实现预期目标和产生潜在影响的方式、算法背后的假设、AIED在准确性和可解释性之间作出的权衡与理由。

在采购阶段，特别需要关注上述证据的解释力度。GDPR要求产品开发企业向利益相关者提供“有意义的解释”(EU，2016)。因此，设计者需要确保这些证据的效度，即确保这些材料能够证明在AIED设计过程中采取了减少偏见的措施，做出了满足各类学习者需求的包容性设计，具备数据收集的正当目的，进行过利益相关者咨询，具有多元化的组成，参与过技术伦理培训，不会强迫学习者或致使其成瘾，确认过产品会对学习者行为产生积极影响等。

此外，在我国本土教育场景下，“有意义的解释”还涉及解释形式、应用情境、面向对象、解释目的等各项具体因素。我国近年来开始大力推广人工智

能课程，但许多地区因师资、设备等条件制约，智能产品对于当地师生来说仍是新兴事物。针对这类用户的特点，学校负责采购的人员应当要求设计者提供更加简明易懂的说明指南。采购者作为学校内部最了解AIED产品的人，还应该承担起更多辅助性的责任，比如积极引导师生熟悉基本操作，教授师生如何在日常生活中应用，着重解释清楚数据收集的正当目的与数据管理机制等，以提前规避伦理争议。

3.使用阶段：提升师生的AI伦理素养水平

如前文所述，过多或过少的信赖都会影响AIED产品的使用情况。师生需要了解AI技术的特点和可能带来的影响，掌握与AI进行有效互动与批判性评估的能力，唯此才能适当地校准他们对于系统的信赖程度。相较于英国在培养AI伦理素养方面的进展，我国目前缺少类似于Apps for Good这一类注重培养学生人工智能伦理素养的专业组织。近年来所出版的中小学人工智能教材中关于人工智能伦理主题的内容也大多作为后置的补充单元，在广度与深度上都有所不足（赵慧臣等，2019）。为提升一线师生恰当使用AIED产品的能力，我国应当进一步推动人工智能伦理教育的发展。

首先，国内高校与研究机构可以加强面向学生的人工智能伦理课程与工作坊开发，从专业角度助力学生AI素养的发展。如美国麻省理工学院媒体实验室（MIT Media Lab）就开发了面向中小学生的“AI与伦理”课程，通过一系列动手实践活动，带领学生学习AI系统的相关技术概念、AI技术的伦理与社会知识。而同样由他们开发的“个人机器人小组”（Personal Robots Group）工作坊，也通过一系列与学生日常生活息息相关的趣味性活动向中小学生传递了数据隐私相关知识（Akgun et al.，2022），这些课程都可以作为我国开发类似教育资源的参考。

其次，开展更多以人工智能为主题的教师专业发展活动，在帮助教师学会在课堂中使用技术的同时，也让教师理解AI只是为教育赋能的一种可选择的方法，并非唯一路径。在“借助智能技术淬炼教师的核心能力”的同时，也让教师警惕对于运用智能技术进行教学判断的过分依赖（汪琼等，2021）。

再次，中小学也应进一步加强与大学、科研机构、人工智能企业的伙伴关系，依托学校的真实教育场景、教师的专业知识与教学智慧，以及其他机构的AI前沿知识，以平等、互惠的合作形式，共同就AI在教育中的应用进行学习与探究。

总体来说，学校与相关机构应协力支持师生更好地理解AI技术及其在教育中的伦理问题，树立起他们对于可信赖AIED的合理预期，进而通过对AIED产品的恰当运用，建立起人与技术的相互信赖，使技术成为增进教育福祉的重大助力。

参考文献：

[1]郭望舒(2009).中西方对于个人空间的比较性研究[D].北京:首都师范大学:31-33.

[2]李应潭(1999).人工智能可信赖性与可信赖算法研究[J].信息与控制,(2):7-13,18.

[3]万海鹏,余胜泉,王琦等(2021).基于学习认知地图的开放学习者模型研究[J].现代教育技术,31(4):97-104.

[4]汪琼,李文超(2021).人工智能助力因材施教:实践误区与对策[J].现代远程教育研究,33(3):12-17,43.

[5]新京报(2019).广雅中学拟采用电子手环被质疑侵犯隐私,校方:还在研究[EB/OL].[2022-04-08].http://www.bjnews.com.cn/edu/2019/03/07/553814.html.

[6]赵慧臣,张娜钰,闫克乐等(2019).高中人工智能教材的特征、反思与改进[J].现代教育技术,29(11):12-18.

[7]AI HLEG (2018). Ethics Guidelines for Trustworthy AI [EB/OL]. [2022-04-08]. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[8]Aiken, R., & Epstein, R. (2000). Ethical Guidelines for AI in Education: Starting a Conversation[J]. International Journal of Artificial Intelligence in Education, 11(2):163-176.

[9]Akgun, S., & Greenhow, C. (2022). Artificial Intelligence in Education: Addressing Ethical Challenges in K-12 Settings[J]. AI and Ethics, 2:431-440.

[10]Baker, T., Smith, L., & Anissa, N. (2019). Educ-AI-tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges[EB/OL]. [2022-06-07]. https://www.nesta.org.uk/report/education-rebooted/.

[11]Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification[C]// Proceedings of the 1st Conference on Fairness, Accountability and Transparency. US: ML Research Press:77-91.

[12]CDEI (2021a). The Roadmap to an Effective AI Assurance Ecosystem[EB/OL]. [2022-05-10]. https://www.gov.uk/government/publications/roadmap-to-an-effective-ai-assurance-ecosystem/roadmap-to-an-effective-ai-assurance-ecosystem.

[13]CDEI (2021b). AI Barometer Part 5 – Education[EB/OL]. [2022-04-08]. https://www.gov.uk/government/publications/ai-barometer-2021/ai-barometer-part-5-education.

[14]Coraggio, G. (2019). First GDPR Fine in Sweden for Illegal Facial Recognition at School[EB/OL]. [2022-05-22]. https://www.gamingtechlaw.com/2019/09/fine-gdpr-sweden.html.

[15]Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way[M]. Switzerland: Springer Cham:50-60.

[16]Edukey (2022). Edukey and GDPR Compliance[EB/OL]. [2022-05-23]. https://www.edukey.co.uk/gdpr-compliance/.

[17]EU (2016). General Data Protection Regulation (GDPR) [EB/OL]. [2022-06-02]. https://gdpr-info.eu/.

[18]FindAMasters (2022). Masters Degrees in Artificial Intelligence, United Kingdom[EB/OL]. [2022-04-30]. https://www.findamasters.com/search/courses.aspx?CID=GB&JD=8&JS=810.

[19]Friedman, B., Kahn, P. H., & Borning, A. (2002). Value Sensitive Design: Theory and Methods[EB/OL]. [2022-04-30]. UW CSE Technical Report. https://research.cs.vt.edu/ns/cs5724papers/6.theoriesofuse.cwaandvsd.friedman.vsd.pdf.

[20]GOV.UK (2016). Artificial Intelligence: Opportunities and Implications for the Future of Decision Making[EB/OL]. [2022-04-08]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf.

[21]GOV.UK (2021). Ethics, Transparency and Accountability Framework for Automated Decision-Making[EB/OL]. [2022-04-08]. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making.

[22]Hall, W., & Pesenti, J. (2017). Growing the Artificial Intelligence Industry in the UK[EB/OL]. [2022-06-06]. https://apo.org.au/node/114781.

[23]Heintzelman, S., & Bathon, J. (2017). Caught on Camera: Special Education Classrooms and Video Surveillance[J]. International Journal of Education Policy & Leadership, 12(6):1-16.

[24]Holmes, W., Porayska-Pomsta, K., & Holstein, K. et al. (2021). Ethics of AI in Education: Towards a Community-Wide Framework[J]. International Journal of Artificial Intelligence in Education, 32:504-526.

[25]House of Lords (2018). AI in the UK: Ready, Willing and Able? [EB/OL]. [2022-06-07]. https://www.politico.eu/wp-content/uploads/2018/04/I-in-the-UK-ReadyWillingAndAble-April-2018.pdf.

[26]IEAIED (2021). The Ethical Framework for AI in Education[EB/OL]. [2022-06-07]. https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/.

[27]Institute for Ethics in AI (2022). Research Themes[EB/OL]. [2022-05-23]. https://www.oxford-aiethics.ox.ac.uk/.

[28]ISO/IEC TR 24028: 2020 (2020). Information Technology-Artificial Intelligence - Overview of Trustworthiness in Artificial Intelligence[EB/OL]. [2022-05-22]. https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:24028:ed-1:v1:en:term:3.47.

[29]Januszewski, A., & Molenda, M. (2013). Educational Technology: A Definition with Commentary[M]. UK: Routledge:10-12.

[30]Kwarteng, K., & Dorries, N. (2021). National AI Strategy[EB/OL]. [2022-06-07]. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf.

[31]Leverhulme Centre for the Future of Intelligence (2022). Master of Studies: AI Ethics & Society[EB/OL]. [2022-05-23]. http://lcfi.ac.uk/master-ai-ethics/.

[32]Luckin, R. (2017). Towards Artificial Intelligence-Based Assessment Systems[J]. Nature Human Behaviour, 1(3):1-3.

[33]Rahim, Z. (2020). Algorithms Can Drive Inequality. Just Look at Britain's School Exam Chaos[EB/OL]. [2022-02-13]. https://www.cnn.com/2020/08/23/tech/algorithms-bias-inequality-intl-gbr/index.html.

[34]Reynolds, M. (2017). UK's Nudge Unit Tests Machine Learning to Rate Schools and GPs[EB/OL]. [2022-04-30]. https://www.wired.co.uk/article/rithms-schools-ofsted-doctors-behavioural-insights.

[35]Roose, K. (2023). Don't Ban ChatGPT in Schools. Teach with It[EB/OL]. [2023-02-07]. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html.

[36]UCAS (2022). Courses in Artificial Intelligence[EB/OL]. [2022-04-30]. https://digital.ucas.com/coursedisplay/results/courses?searchTerm=artificial%20intelligence&destination=Undergraduate&distanceFromPostcode=&studyYear=2022&sort=MostRelevant.

[37]Vincent-Lancrin, S., & Van der Vlies, R. (2020). Trustworthy Artificial Intelligence (AI) in Education: Promises and Challenges[EB/OL]. [2022-06-07]. https://www.oecd-ilibrary.org/education/tificial-intelligence-ai-in-education_a6c90fa9-en.

收稿日期 2023-03-05 责任编辑 刘选

Path and Enlightenment of Trustworthy Artificial Intelligence in Education: A Case Study of Typical Initiatives in the United Kingdom

SHEN Yuan, HU Mengyuan, FAN Yizhou, WANG Qiong

Abstract: Artificial intelligence (AI) has shown great potential in assisting teaching and learning. However, its opacity, unpredictability, and improper use have sparked controversies. Building trustworthy AI in education (AIED) has

become an international consensus. Trustworthy AIED emphasizes a "people-centered" value, with reliable and secure AI technology as the underlying foundation and ethical understanding and usage as critical safeguards. Recognizing the complexity of educational contexts, it rejects the practice of mechanically applying general AI ethics principles or professional ethics principles from other AI application domains to the field of education. The United Kingdom, as a pioneer in AI ethics governance, has accumulated rich experience in constructing trustworthy AI in education, particularly evident in *The Roadmap to an Effective AI Assurance Ecosystem*. This roadmap starts from three significant phases: technology development, procurement and utilization. It adopts specific strategies such as ethical design, ethical verification, and AI literacy cultivation to establish a guarantee system for trustworthy AI in education. China should draw lessons from the UK's specific initiatives in AI in education. During the development phase, it is essential to conduct value-sensitive design of intelligent education products that align with local values. In the procurement phase, meaningful explanations should be provided for the selection of intelligent education products. During the utilization phase, efforts should be made to cultivate and enhance AI literacy among teachers and students. By doing so, China can promote the local development of trustworthy AI in education by involving technology developers, product procurers and frontline users.

Keywords: Artificial Intelligence in Education; Trustworthy Artificial Intelligence; Ethical Design; UK Initiatives

(上接第64页)

The Legislative Dilemma and Legal System Construction of Elderly Education in China Under the Background of Aging Population

WANG Shaoyuan, LIU Bowei

Abstract: With the acceleration of China's aging population process, how to integrate a positive aging perspective into the entire process of economic and social development, integrate elderly education into the legal track, and plan the construction of the legal system for elderly education with a systematic thinking of the rule of law has become an important topic of the times. From the perspective of the legislative process and basic characteristics of elderly education in China, there are mainly difficulties such as a weak legal system for elderly education, insufficient supply of legal systems, vague legislative value concepts, and legislative regulation dominated by policy and law adjustments. Only by resolving the systemic defects in the overall structure and function of the elderly education law can we better ensure the full realization of the right of the elderly to education. The systematic construction of legislation on elderly education in China in the future can take three basic paths. Firstly, based on the value concepts of natural rights, inherent rights, and exclusive rights, the internal value system of elderly education law should be constructed; Secondly, taking the normative system, implementation system, supervision system, and guarantee system as the formal rationality, the overall structure of the elderly education law should be improved; Thirdly, with the goal of strengthening the function and effectiveness of the system, the Basic Law on Elderly Education in China should be formulated, and hence to form a elderly education legal system with a progressive logic, guided by the Constitution, led by the Education Law, based on the Law on the Protection of the Rights and Interests of the Elderly, with the Special Law on Elderly Education as the core and supplemented by local regulations.

Keywords: Education for the Elderly; Elderly Education Legislation; Legal Systematization; The Right to Education; Population Aging