DOI:10.13811/j.cnki.eer.2019.07.003

智能教育时代下人工智能伦理的内涵与建构原则

杜 静，黄荣怀，李政璇，周 伟，田 阳

（北京师范大学 智慧学习研究院，北京 100875）

[摘 要] 在智能教育时代，人机如何共处是人工智能伦理建构的关键。文章首先从技术悖论视角，厘清当前人工智能应用于教育在技术滥用、数据泄露、智能教学机器的身份与权力边界等方面存在的伦理挑战与困境；其次，利用内容分析法，结合多国与国际组织政策文件，对人工智能伦理相关的伦理要素进行分析与抽取，发现政府、高校、国际组织文件中多次提到的价值、人类利益、安全、隐私、责任等关键要素；最后，基于人机共处的考量，结合人工智能在教育领域的应用现状和伦理关键要素，归纳分析出智能教育伦理需遵循的原则，包括问责原则、隐私原则、平等原则、透明原则、不伤害原则、非独立原则、预警原则与稳定原则。

[关键词] 智能教育；人工智能；机器伦理；国际政策；原则；APETHICS 模型

[中图分类号] G434 [文献标志码] A

[作者简介] 杜静(1989—)，女，湖北武汉人。博士研究生，主要从事适应性学习支持服务研究。E-mail：dujing@mail.bnu.edu.cn。黄荣怀为通讯作者，E-mail：huangrh@bnu.edu.cn。

科学技术的发展与伦理密不可分。人工智能技术带来便捷的同时，也衍生出了复杂的伦理、法律和安全问题，需要相关的规范制定者和执行者提前布局，审慎应对。在教育领域，一方面，人们期待人工智能进一步促进教育的发展，如支持个性化学习、提供教学适切服务等，且伴随着人工智能进入高中课本，中国的人工智能教育已经正式进入基础教育阶段。在人工智能技术领先的美国，从 2018 年秋季开始，匹兹堡蒙托学区也启动了全美第一个人工智能公立学校项目。另一方面，随着智能教育应用的广泛开展，教育工作者尝试通过“先进技术”解决教学难题，在这过程中隐含一系列伦理道德的问题，如知识产权保护、隐私泄露、学术不端等，这使人们开始反思不当使用“先进技术”会造成什么负面影响，进而衍生出“人工智能进入教育后，人机如何共存”的社会问题。国际上已经开始关注人工智能教育应用过程中存在的伦理问题，然而当前关于此领域的研究仍处在探索早期。为此，本文从技术悖论现象的视角，深度剖析人工智能应用于教育的伦理困境，结合国际案例剖析了人工智能教育应用面对的伦理挑战，并基于国内外人工智能领域伦理问题的国家政策，提出智能教育伦理建构原则，以期为人工智能教育应用和发展提供伦理依据和启发性思考。

一、智能时代的伦理议题

自 1956 年，在美国达特茅斯(Dartmouth)会议上确定人工智能(Artificial Intelligence，简称 AI)这一术语，经过 60 多年的发展，该技术已在图像识别、机器视觉、自动驾驶、安防等领域的发展取得较大进展。在发展的过程中由于缺乏关乎伦理道德的考量，引发了一系列的问题，这使得越来越多的行业专家开始反思智能时代人机如何共处。

伦理一词在中国最早出现在《礼记·乐记》：“凡音者，生于人心者也；乐者，通伦理者也”，译为一切音乐

基金项目：教育部科技司委托项目“中国智能教育推进路径研究”(项目编号：教技司[2018]477 号)；江苏高校哲学社会科学研究基金项目“混合学习的学习交互性社交研究”(项目编号：2017SJB0573)

都产生于人的内心,乐与事物的伦理相同。乐与礼的关系,在《礼记》中指的也就是文艺与社会伦理道德的关系[1]。剑桥大学哲学词典(The Cambridge Dictionary of Philosophy)指出,“伦理”(Ethics)一词,广义上指“道德”,狭义上常被用来表示某一特定传统、群体或个人的道德原则[2]。伦理学试图通过定义诸如善与恶、对与错、美德与罪恶、正义与犯罪等概念来解决人类道德问题,常常与道德心理学、描述性伦理、价值理论等领域的研究有关[3]。综上所述,多数研究主要从政治、道德等方面探讨伦理的内涵,集中于人、社会、自然间伦理关系的阐述。因科技的进步与发展,智能机器与人类生活愈发密切,我们需要进一步审视人与智能机器的伦理关系。由此归纳在智能时代伦理的应有内涵,指在一定社会历史条件下,为解决人、社会、自然与智能机器的和谐共处问题,通过制定相关原则、标准和制度,促进人、社会、自然和智能机器的和谐共处,涉及道德、社会制度、法律等领域。伦理研究的本质应是一个不断完善的道德实践过程。

二、技术悖论视角下人工智能应用于教育的伦理困境审视

辩证唯物主义认识论是技术辩论的思想来源,辩证唯物主义告诉我们事物是普遍联系与存在矛盾的,看待事物的发展需要把握好两点论[4],即要同时关注事物发展过程中的正面与反面。早在 19 世纪中期,马克思就认识到技术悖论对人类社会产生的双面影响,他指出:“在我们这个时代,每一种事物都包含有自己的反面。我们看到,机器具有减少人类劳动和使劳动更有成效的神奇力量,然而却引起了饥饿和过度的疲劳。技术的胜利,似乎是以道德的败坏为代价换来的。”[5]人工智能技术的发展亦符合技术悖论现象,我们必须正视人工智能为教育领域带来的种种福利,如支持个性化学习、提供学习过程适切服务、提升学业测评精准性、助力教师角色变化等[6],也不可否认在人工智能技术应用过程中由于伦理制度的缺失、公众道德素质与文化素养的不足以及政策法规的滞后与不完善等引发的伦理问题。

(一)技术滥用引发的不端行为

或许人工智能技术并不像人类想象的那样美好,在应用的过程中仍旧潜藏着一定的危险。在 Nature 评论上,2018 年 4 月 25 日 17 位来自杜克大学、斯坦福大学、哈佛大学等诸多高校的教授和科学家联合发表文章,指出现在需要对培养人脑组织的行为作出伦理反思,并提出亟须建立伦理道德框架以应对这一难题[7]。剑桥大学、牛津大学与耶鲁大学的学者认为,在未来 5 到 10 年人工智能系统可能催生新型网络犯罪、实体攻击和政治颠覆,设计这项技术的人需要做更多的事情来降低技术被误用的风险,此外,政府必须考虑制定新的法律[8]。除此之外,技术滥用还有可能引发学术不端现象。Ikanth & Asmatulu 指出,智能手机是如今最为普遍的作弊工具,大约 70%的学生承认他们在考试、作业、团队任务、报告和论文的写作过程中使用各种高科技设备,比如 iPad 等进行过作弊行为[9]。从社会现象来看,由于目前社会对伦理问题的认知不足以及相应规范准则的缺失,设计者在开发人工智能教育产品时并不能准确预知可能存在的风险,人工智能与教育的深度融合趋势明显,必须考虑更深层的伦理问题,从而使人工智能教育应用产品的设计目标与更好地服务学习者的初衷相符。

(二)数据泄露引发的隐私担忧

如今教学系统功能愈发智能和丰富,不仅仅可以通过指纹、人脸、声音等生理特征识别用户身份,还能够搜集和记录环境信息。Inayay 开发出可实时监控学生课堂活动的定位系统和学习行为可视化系统,并已将该系统应用于 132 名师生[10]。课堂环境中的温度、湿度和二氧化碳浓度严重影响了教师和学生的学习效率,并影响师生的身体健康,为此有研究者设计了一套基于蓝牙的课堂环境实时监测系统[11]。此外,还可以利用智能运动设备,如智能手环、智能肺活量等测评工具,深度采集学生健康数据,从而发现学生在体质、运动技能、健康程度等方面的问题[12]。智能系统掌握了大量的个人行为信息,如果缺乏隐私保护,就可能造成数据泄露。如果这些数据使用得当,可以提升学习服务的支持效果,但如果某用户出于某些目的非法使用行为信息,会造成隐私侵犯,甚至是违法事件。因此,设计智能系统时需要纳入隐私保护功能。

(三)智能教学机器的身份与权力边界

智能教学机器为一对一学习支持服务的实现提供了可能,可以针对不同的学生采用不同的教学方法,激发学习兴趣,从而提升学习者的学习表现。如果长时间让学生和智能教学机器待在一起,这些机器是否会取代教师的身份?人机交互过多是否会使学生出现社交障碍?欧盟已经在考虑要不要赋予智能机器人“电子人”的法律人格[13],这意味着机器人也具备了合法身份,很难确定机器人享有的权利以及能否履行义务,并对其行为负责。随着智能教学机器的功能越来越强大,它在教学的过程中到底应该扮演一个怎样的角色。此外,在机器被赋予决策权后,智能教学机器在

何种情境下辅助学生学习才能够帮助学生达到更好的学习效果。在平板上记笔记的学生在概念性问题表现上比普通书写方式记笔记的学生差,在平板上记笔记的学生虽然花了更长时间记笔记且覆盖了更多内容,但因为借助科技而被动不走心的记忆几乎把这些好处抵消掉了[9]。这些研究表明,技术不如教师具有亲和力,在某些时候技术的使用会降低学习者的学习体验。

三、国际人工智能伦理研究概况

一些国家相继成立了人工智能伦理研究的各类组织和机构探讨人工智能引发的伦理问题,如科学家组织、学术团体和协会、高校研发机构,还有国家层面的专业性监管组织。这些国际政策将为人工智能教育应用的伦理模型构建提供理论依据。

(一)多国人工智能伦理政策研究动态

研究样本选取包括美国、英国、德国、中国在内二十余国的人工智能政策文件,检索截止时间为 2016 年 1 月至 2018 年 8 月,搜索源为各国政府网站,搜索关键词为 Artificial Intelligence 或者 AI。再对下载的文件进行二次检索,关键词为 ethics 或 ethic。如遇有的国家直到 2018 年 5 月也无相关文件公开发布,则将最新一次召开的人工智能会议记录下来,以此作为该国人工智能研究进展,如突尼斯对人工智能的研究起步较晚,于 2018 年 4 月刚召开完第一次人工智能会议,在本次统计中同样进行记录。在对各国政策文件进行了内容抽取后,绘制了人工智能国家政策概览,如图 1 所示。

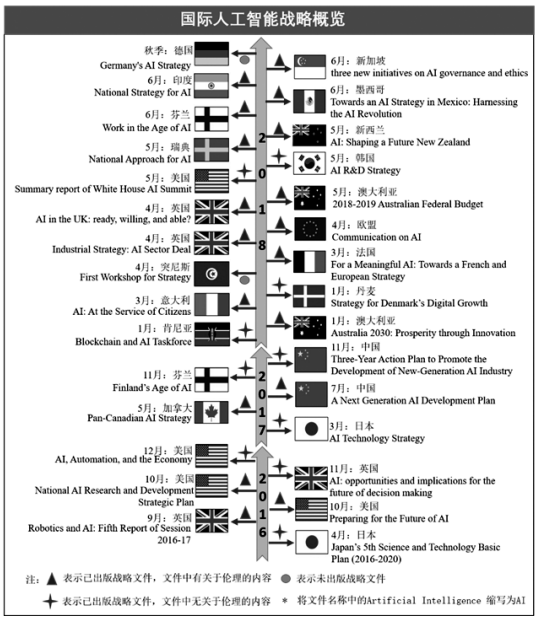

图 1　多国人工智能战略研究概览

各国对人工智能的关注持续升温,在该领域的战略研究主要集中在科学研究、人才培养、技能培训、伦理与安全、标准与法规、基础设施建设等方面。美国、英国、日本对人工智能的研究关注较早,政府已出版多份人工智能战略文件。英国已出版四份战略文件来阐述未来将会如何规范机器人技术与人工智能系统的发展,以及如何应对其发展带来的伦理道德、法律及社会问题,包括《人工智能:未来决策制定的机遇与影响》(Artificial Intelligence: Opportunities and Implications for the Future of Decision Making)、《机器人技术和人工智能》(Robotics and AI: Fifth Report of Session 2016—2017)、《人工智能行业新政》(Industrial Strategy: AI Sector Deal)和《英国人工智能调查报告》(AI in the UK:Ready,Willing,and Able)。美国于 2016 年连续发布三份报告,为美国的 AI 战略奠定了基础,包括《为人工智能的未来做好准备》(Preparing for the Future of Artificial Intelligence)、《国家人工智能研究和发展战略计划》(National AI Research and Development Strategic Plan)、《人工智能、自动化与经济报告》(AI,Automation,and the Economy),之后于 2018 年 5 月,白宫邀请了业界、学术界和部分政府代表参加了一场人工智能峰会,并发布人工智能业界峰会总结(Summary Report of White House AI Summit)。而肯尼亚、德国对人工智能战略研究起步较晚,德国计划在 2018 年秋季发布国家人工智能战略,但目前还不清楚该战略将关注哪些方面。肯尼亚政府于 2018 年 1 月宣布将成立一个新的 11 人特别工作组,该小组的任务是制定鼓励开发和采用如区块链和人工智能等新技术的国家战略[14]。在芬兰、美国、英国等国的战略文件中,已明确提出将关注人工智能引发的伦理问题。

从世界各国政策文件看,发达国家对人工智能各方面关注较为全面,尤其是将伦理问题上升到了一定的高度,而发展中国家更多的是关注人工智能的技术应用和未来的商业价值等。由此,发达国家的人工智能政策虽然值得借鉴,但是在伦理研究方面各国仍处于早期状态,这对人工智能发展也有一定的制约作用。

(二)国际组织人工智能伦理研究动态

除了国家层面关注人工智能的伦理问题研究,一些传统的国际性组织和高校也开始关注人工智能治理的问题,见表 1。2016 年 12 月和 2017 年 12 月,标准制定组织 IEEE(Institute of Electrical and Electronic Engineers,美国电气和电子工程师协会)发布《合伦理设计(第一版)》和《合伦理设计(第二版)》,并首次在

报告中考虑人类生存幸福感(Well-being)的问题,由IEEE下属各委员会共同完成的文件为人工智能的全球治理提供了重要参考。UNESCO (United Nations Educational,Scientific,and Cultural Organization,联合国教科文组织)和COMEST(World Commission on the Ethics of Scientific Knowledge and Technology,世界科学知识与技术伦理委员会)多年来连续多次联合发布报告,就机器人应用过程中的伦理问题展开了讨论,包括伦理困境、如何保证创新是符合伦理的,如《COMEST机器人道德报告》[15],这对世界各国的人工智能监管具有重要指导意义。此外,EURON (European Robotics Research Network,欧洲机器人研究网络)也曾对下一代人形机器人发展过程中涉及的技术二重性等伦理问题进行简要概述[16]。斯坦福大学的"人工智能百年研究项目"计划"针对人工智能在自动化、国家安全、心理学、道德、法律、隐私、民主以及其他问题上所能产生的影响,定期开展一系列的研究。"该项目的第一份研究报告《人工智能2030生活愿景》已经于2016年9月发表。卡内基梅隆等多所大学的研究人员联合发布《美国机器人路线图》,以应对人工智能对伦理和安全带来的挑战。来自行业和学术界包括牛津大学、剑桥大学、新美国安全中心(CNAS, Center for A New American Security)的人工智能专家撰写《恶意使用人工智能风险防范:预测、预防和消减措施》,调查了人工智能恶意使用的潜在安全威胁,并提出了更好的预测、预防和减轻这些威胁的方法[8]。

多国政策主要集中于人工智能对工业生产、经济、文化、社会、国防安全等方面带来的颠覆性影响,但又有不同的侧重点。美国与中国强调人工智能技术的发展,如美国关注人工智能的研发生态、人才培训战略、行业应用与国防安全,中国关注研发、工业化、人才发展、教育和职业培训、标准制定和法规、道德规范与安全等各个方面的战略;日本和印度关注人工智能在具体领域的应用,如日本关注人工智能在机器人、汽车等领域的落地,印度聚焦于人工智能在健康护理、农业、教育、智慧城市和基础设施与智能交通五大领域的应用;欧盟对人工智能伦理问题的研究更为关注,正在研究人工智能伦理框架。这些文件中有关伦理的研究为厘清人工智能融入教育所带来的伦理问题提供借鉴和参考。

四、面向智能教育的人工智能伦理建构原则

现代科学技术在给教育带来便利的同时,也使广大的教育工作者和学习者不得不置身于伦理困境之中。国际政策关于人工智能伦理问题的研究为智能教育伦理建构原则的提出奠定了坚实的政策基础和前瞻性认识。

(一)有道德人工智能的特征

科学研究无法脱离特定历史时期对技术研究的价值取向,每一次的科学研究都蕴含着研究主体对技术手段、研究方向和研究方法的选择,科学技术产生的社会作用越大,与伦理的关系也将越来越紧密。人工智能技术的快速发展,也必然引发社会的担忧。有学者曾经提出,判定人工智能程序是否是符合伦理道德的初步设想,见公式(1)[17]。该学者认为,有道德的人工智能研究需要从人工智能技术的发展逐步过渡到有道德地使用人工智能,但该学者并未明确指出怎样的用户行为属于有道德的使用人工智能。

$$\text{AI research is ethical} \equiv \text{AI research} \rightarrow \text{ethical AI} \quad \text{公式(1)}$$

那么人工智能技术的发展是否有道德规范可循?早在1942年,艾萨克·阿西莫夫(Isaac Asimov)提出了以"不得危害人类"为核心的"机器人三原则"(The Three Laws of Robotics)(包括:机器人不得伤害人类

表1 国际机构和高校人工智能伦理研究概览

机构或高校名称	报告名称(英文)	报告名称(中文)	时间
牛津大学、剑桥大学、新美国安全中心等机构联合发布	The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	恶意使用人工智能风险防范:预测、预防和消减措施	2018.2
IEEE	Ethically Aligned Design Version 2	合伦理设计(第二版)	2017.12
UNESCO & COMEST	Report of COMEST on Robotics Ethics	COMEST机器人道德报告	2017.9
Future of Life Institute(生命研究院)	Asilomar AI Principles	阿西洛马人工智能原则	2017.2
IEEE	Ethically Aligned Design Version 1	合伦理设计(第一版)	2016.12
卡内基梅隆大学等多所大学联合发布	From Internet to Robotics: A Roadmap for US Robotics	美国机器人路线图	2016.11
斯坦福大学	Artificial Intelligence and Life in 2030	2030人工智能与生活前景	2016.9
EURON	Roboethics Roadmap	机器人伦理路线图	2006.3

个体，或者目睹人类个体将遭受危险而袖手不管；机器人必须服从人给予它的命令，当该命令与第一定律冲突时例外；机器人在不违反第一、第二定律的情况下要尽可能保护自己的生存)[18]。三定律针对机器人和人两大主体规定了机器人的行为规则，但该定律并未给出有关于人类和机器人的明确的、统一的定义。在 1983 年，他又在此基础上添加了零原则，即机器人必须保护人类的整体利益不受伤害，其他三条定律都是在这一前提下才能成立[19]。新“机器人三原则”被认为是人类对人工智能伦理道德反思的开端。

(二)人工智能伦理关键要素分析

在搜集国际人工智能政策文件的基础上，利用内容分析法对相关文件中的伦理要素进行了提取，见表 2。政府、高校、国际组织均尝试对人工智能的伦理问题进行探讨，从伦理相关要素看，均对人工智能立法、安全、责任认定等有着一致性，但也有差异问题。政府希望通过制定法律保护公民的权益，如个人隐私、明确责任主体；国际组织更加强调人工智能社会价值的探讨，如人类利益、机器意识等关键词多次出现；高校则结合实际问题提出了应对人工智能风险问题。除此以外，有研究者指出，引入智能机器带来的全球性的社会和伦理问题可总结为技术的二元性、技术产品的人格化、技术沉迷、数字鸿沟、数字资源公平享有权、技术对全球财富和权力的影响、技术对环境的影响 [20]。工程公司美国理事会伦理守则(American Council of Engineering Companies Ethical Guidelines)还指出，安全性和可靠性也是人造产品需要考虑的道德因素[21]，同时，还要注意人权[22]、责任权、公正权、正义、非歧视原则、知情权、社会责任、平等权。

(三)智能教育伦理的建构原则

由于人工智能的交叉性，人工智能可与其他应用伦理学共享问题和解决方案，如计算机伦理、信息伦理、生物伦理学、技术伦理和神经伦理学。计算机和信息伦理学已发展出了名为 PAPA 的道德准则，即隐私权(Privacy)、准确性(Accuracy)、所有权(Property)、易获得性(Accessibility)[23]。为此，在参照 PAPA 模型的基础上，基于对人工智能伦理的相关要素分析，立足人工智能在教育领域的应用现状，从人机共存角度，抽取与学生发展密切相关的要素，涉及明确责任主体、保护人类隐私、不偏见不歧视、决策透明化、保护人类利益不受侵害、提前预警危险行为、系统稳定可控等多个方面，最终归纳出面向智能教育的人工智能伦理建构模型，即问责原则 (Principle of Accountability)、隐私原则(Principle of Privacy)、平等原则 (Principle of Equality)、透明原则 (Principle of Transparency)、不伤害原则(Principle of Noharm)、身

表 2　　伦理的相关要素

类别	来源	报告名称(中文)	伦理相关要素
政府	美国	人工智能行动法案	借助算法消除歧视；机器的责任
	美国	为人工智能的未来做好准备	公平、安全与伦理；人工智能的安全与可预测性
	英国	人工智能：未来决策制定的机遇与影响	个人隐私；知情权；由人工智能进行决策的问责概念和机制；算法偏差导致的偏见风险
	中国	新一代人工智能发展规划	追溯和问责；人工智能法律主体及相关责任；人工智能产品设计人员的道德规范和行为准则；人工智能的法律法规
国际组织	IEEE	合伦理设计(第二版)	人类权力；优先考虑人类幸福感；问责原则；透明原则；技术滥用意识
	UNESCO & COMEST	COMEST 机器人道德报告	人权；“不伤害”原则；问责原则；善行原则；正义原则；自主权；保护隐私
	生命研究院	阿西洛马人工智能原则	安全性；故障透明性；司法透明性；责任；价值归属；人类价值观；个人隐私；自由与隐私；分享礼仪；共同繁荣；人类控制；非颠覆
	EURON	机器人伦理路线图	技术二重性；机器格化；人机关系的人性化；技术沉迷；数字鸿沟；技术资源获取的非平等性；技术对全球权力与财富分配的影响；技术对环境的影响
	IEEE	合伦理设计(第一版)	保护人类利益原则；问责原则；透明原则；机器人的教育与意识
高校	卡内基梅隆大学等多所大学联合发布	美国机器人路线图	安全；可靠性；隐私
	斯坦福大学	2030 人工智能与生活前景	数据安全与隐私；完善人工智能法律与政策体系
	牛津大学、剑桥大学、新美国安全中心等机构联合发布	恶意使用人工智能风险防范：预测、预防和消减措施	黑客攻击；深度学习的“黑箱”决策过程；相关用户的行为约束

份认同原则(Principle of Identity)、预警原则(Principle of Precaution)、稳定原则(Principle of Stability),概括为“APETHICS”模型,如图 2 所示。需要说明的是,取单词 Caution 首字母 C 代表预警原则,其余均取单词首字母表示各原则, 缩写字母 A 在本模型中代表问责原则, 也可以看作为 Artificial Intelligence 中 Artificial 的首字母,P 在本模型中代表隐私原则,也可以看作为“+”(Plus),其余的取单词首字母,恰好为伦理“Ethics”一词,合并后为 A(rtificial Intelligence) & P(lus) & Ethics。以下将对原则进行具体解读:

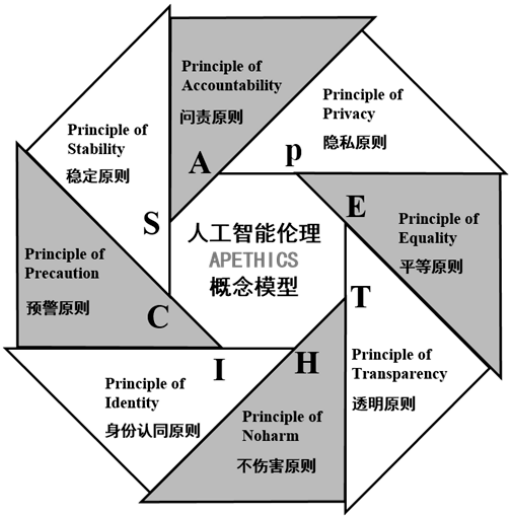

图 2　智能教育伦理“APETHICS”模型

1. 问责原则(Principle of Accountability)

“问责原则”主要是指明确责任主体,建立具体的法律明确说明为什么以及采取何种方式让智能系统的设计者和部署者承担应有的义务与责任。问责是面向各类行为主体建立的多层责任制度。然而,分清楚责任人是极具挑战性的,例如,在自动驾驶汽车独立作出智能决策导致伤害发生的情况下,谁应当为自动驾驶汽车发生的故障和事故负责的问题,是应该由司机、自动汽车生产者或是人工智能系统设计公司来承担法律责任?在这里还难以作出明确的判断。《为人工智能的未来做好准备》提出了一般性的应对方法,强调基于风险评估和成本收益原则决定是否对人工智能技术的研发与应用施以监管负担[24]。在 UNESCO 和 COMEST 联合发布的《Report of COMEST on Robotics Ethics》中尝试明确责任主体[15],见表 3。在追责政策制定不完善、主体责任不明确、监督责任不到位、伤害人类行为多发频发等问题发生时起到一定的缓解作用,以问责倒逼责任落实,落实科学家、设计者、政策制定者、使用者的各级责任,争取做到“失责必问、问责必严”。

2. 隐私原则(Principle of Privacy)

《辞海》将隐私定义为公民依法享有的不公开与其私人生活有关的事实和秘密的权利。《世界人权宣言》曾经明确规定,“任何人的私生活、家庭、住宅和通信不得任意干涉,他的荣誉和名誉不得加以攻击”[25]。在人工智能时代,“隐私原则”更强调人们应该有权利存取、管理和控制智能机器产生的数据,以确保机器不会向任何未经授权的个人或企业提供用户信息。如今,我国有关如何在人工智能时代保护学习者隐私不受侵害的法律还不健全,关于何为侵犯隐私、何种行为侵犯隐私、如果侵犯如何处罚的相关法律仍缺失,倘若学习者隐私受到侵害也无法找到合适的解决途径。国外的人工智能研究人员已经在提倡如何在深度学习过程中保护个人隐私[26]。人工智能时代为学习者隐私保驾护航,将是促进社会和谐发展与长足进步的不可或缺因素。

3. 平等原则(Principle of Equality)

“平等原则”指杜绝因算法偏差导致的算法歧视现象。基于种族、生活方式或居住地的分析是对个人常态数据的分析,该种算法风险可以避免,而“算法歧视” 的实质是用过去的数据预测用户未来的表现,如果用过去的不准确或者有偏见的数据来训练算法,得出的结果肯定也是有偏见的。不少研究者坚持,数字不会说谎,可以代表客观事实,甚至可以公平地为社会服务。但是,已有研究发现,建立在数据基础之上的算法系统也会犯错、带有偏见,而且比以往出现的数字歧视更加隐秘。亚马逊尝试了价格歧视,这意味着不同的消费者购买同一件商品的价格不同[27]。为消除

表 3　　机器决策、人类行为与责任主体关系

由机器作出的决策类型	人类参与的行为	技术的定位	责任方	需要遵守的规则
根据预定设计严格标准,在有限的范围内作出选择	在法律框架内实施标准	只涉及机器:确定的算法或机器人	制作者	合法的(标准;本国或国际的法律)
根据预先设定的政策,在一系列的预选中灵活选择	将决策权授予机器人	只涉及机器:基于 AI 的算法或具有认知能力的机器人	设计者、制造商、卖家、用户	为工程师和用户设定的用户守则;预警原则
通过人机交互作出选择	人类控制机器的选择	机器引发严重有害或致死的行为时,人类有能力控制机器	全人类	伦理道德

该风险，数据人员应该确定数据中存在的某种偏差，并采用相应的策略评估该偏差带来的影响。

4. 透明原则(Principle of Transparency)

“透明原则”指明确说明使用了哪些算法、哪些参数、哪些数据实现了什么目的，机器的运作规则和算法让所有人都能够明白。机器需要了解学习者的行为以作出决策，所有人包括学习者也必须了解机器是如何看待自己和分析自己的处理过程。如果学生个人画像不正确，却无法纠正，该怎么办？当系统搜集学生的信息，却得出错误的结论，该怎么办？目前被广泛讨论的是深度学习的“黑箱”决策过程，它拥有更多的自主决策权，许多研究者试图打开这个“黑箱”。如利用反事实调查的人工智能神经科学工具理解这种黑箱[28]。这种迫切性不仅出自科学[29]。美国国防高级项目研究计划局（The Defense Advanced Research Projects Agency，简称 DARPA）为名为“可解释 AI”（Explainable AI）的新计划投入了 7000 万美元，旨在对支撑无人机与情报挖掘行动的深度学习作出解释[30]。伦理和设计往往息息相关，人工智能开发必须警惕社会和文化偏见，确保研究具有普适性的算法，消除算法歧视存在的可能空间。

5. 不伤害原则(Principle of Noharm)

“不伤害原则”也可以理解为 Principle of Do Not Harm，指必须阻拦机器的无意识行为对人类造成的伤害，任何情况下，不区别对待文化、种族、民族、经济地位、年龄、区域等因素，都要维护人类的权益，不得侵害人类权益。利弊相生，人工智能技术的二重性在给人类的生活和生产带来便利的同时，也会因技术运用不当给人类带来毁灭性的灾难。Google 团队对 DeepMind 人工智能系统进行行为测试，通过设计“搜集水果游戏”和“群狼狩猎游戏”来模拟当多个 DeepMind 人工智能系统具有相似或冲突的目标时，是内斗还是合作。初步的研究结果表明，机器人以及人工智能系统具有“杀手潜能”，它们并不会自动地将人类的利益放在“心”上[31]，设计算法的技术人员需要考虑人性的因素。

6. 身份认同原则(Principle of Identity)

“身份认同原则”指明确智能机器的“社会身份”，从而规范其权利与义务。例如，是否该赋予机器人公民身份与人的权利。智能机器是世界物质组成的一部分，非独立存在，智能机器的身份问题日益重要。人工智能融入教育，已经开始影响现存的人类社会结构，人—社会—自然三元社会正在逐渐地向人—社会—自然—智能机器四元社会发展。无论是先秦的“天人合一”论，还是宋明的“万物一体”论，都强调人与人、人与物、人与社会以及人与自然的整体性、和谐性、统一性关系[32]。在人工智能时代，需要强调的是，人与人、人与智能机器、人与社会、人与自然、机器与自然的整体性、和谐性与统一性，为此需要确立人工智能在教育中的定位，以保障技术、教师、学生、环境之间的平衡。

7. 预警原则(Principle of Precaution)

“预警原则”强调当机器出现危害人类的行为时，采取行动避免伤害，因此，人类需要对机器的行为进行监管，并开发相应的预警技术。危害人类和环境的行为包括：威胁人类的生命；严重的或者不可逆的伤害人类权益的行为[33]。谷歌和牛津大学联合发起的人工智能系统“自我毁灭装置”（Kill Switch）研究项目，这个装置能够让人工智能系统处于不间断的人类监管干预之下。通过算法和功能，让人工智能系统摆脱不良行为，杜绝危害人类行为事件的发生。

8. 稳定原则(Principle of Stability)

“稳定原则”指系统算法稳定且一致，系统不出现不必要的行为或者功能上的非一致性和异常结果。人工智能系统应确保运行可靠安全，避免在不可预见的情况下造成伤害，或者被人恶意操纵，实施有害行为。如果我们希望利用技术来满足人类的需要，人类就必须更深入地理解和尊重机器，从而更好地发挥智能机器的作用。需要指出的是，对于知识创造性的工作，比如医疗、教育培训中具有高度“不确定性”的工作，其承担者则非人类莫属。

五、总结与展望

将人工智能广泛而有效地应用于教学，是未来学校教育发展的必然趋势，具有更强的灵活性、包容性、个性化的人工智能技术可以助力重塑学校教育，进一步提升教学的效果、效率和效益，以适应现代信息化、数字化、智能化的学习型和创新型社会的需要，同时，为教育 2030 的实现提供强有力的智力支持。在预见人工智能技术巨大潜能的同时，也不可忽略道德规范、产品标准和安全规范的社会呼吁和学术研究。对待人工智能，我们仍旧应当保持“乐观>焦虑”的态度，让人工智能始终服务于我们的工作和生活，形成造福社会、推动发展、助力改革的无尽活力。本文深度剖析人工智能教育应用的伦理困境，试图通过人工智能教育应用伦理的“APETHICS”模型与建构原则，厘清人工智能教育应用的价值观念、社会责任。在今后的研究中，需进一步细化概念模型中包含的原则。

[参考文献]

[1] 张恩普.《礼记·乐记》文学批评思想探讨[J]. 古籍整理研究学刊,2006(1):35-37.

[2] AUDI R, AUDI P. The cambridge dictionary of philosophy[M]. United Kindom: Cambridge University Press,1995.

[3] Wikipedia. Ethics[EB/OL]. (2001-12-19) [2018-07-13]. https://en.wikipedia.org/wiki/Ethics#cite_note-2.

[4] 张源东.社会技术与社会技术悖论浅析[J].学理论,2018(4):79-81.

[5] 马克思,恩格斯.马克思恩格斯选集:第 1 卷[M].北京:人民出版社, 1972.

[6] 刘德建,杜静,姜男,等.新一轮人工智能融入学校教育的趋势分析[J]. 开放教育研究,2018,24(4):33-42.

[7] FARAHANY N A, GREELY H T, HYMAN S, et al. The ethics of experimenting with human brain tissue [J]. Nature, 2018, 556 (7702):429-432.

[8] BRUNDAGE M, AVIN S, CLARK J, et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation[EB/OL]. (2018-02-01) [2018-07-13]. https://maliciousaireport.com/.

[9] IKANTH M, ASMATULU R. Modern cheating techniques, their adverse effects on engineering education and preventions [J]. International journal of mechanical engineering education, 2016, 42(2):129-140.

[10] INAYAT I. Real-time student visualization system in classroom using RFID based on UTAUT model [J]. International journal of information and teaching technology, 2017, 34(3):274-288.

[11] PENG K, CHEN J. A classroom environment monitoring system based on bluetooth [J]. Electronic science & technology, 2017,30 (2):123-129.

[12] 余胜泉. 人工智能教师的未来角色[J]. 开放教育研究,2018,24(1):16-28.

[13] GUGARDIAN. Give robots ´personhood´ status, EU committee argues[EB/OL]. (2017-01-12) [2018-07-13]. https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues.

[14] WALLSTREET K. Kenya govt sets blockchain & artificial intelligence taskforce! [EB/OL] (2018-01-16) [2018-07-13] https://kenyanwallstreet.com/kenya-govt-sets-blockchain-artificial-intelligence-taskforce/.

[15] UNESCO & COMEST. Report of COMEST on robotics ethics [R]. Praris: UNESCO, 2017.

[16] VERUGGIO G. The EURON roboethics roadmap: HR 2007: proceedings of international conference on humanoid robots, pittsburgh, 29 november - 1 december 2007[C]. Pennsylvania: IEEE, 2007.

[17] KORB K B. Ethics of AI[EB/OL]. (2007-01-01) [2018-07-13]. https://www.researchgate.net/publication/247934914_Ethics_of_AI.

[18] MORAN M. Three laws of robotics and surgery[J]. Journal of endourology, 2008, 22(22): 1557-1560.

[19] COECKELBERGH M. Moral appearances: emotions, robots, and human morality [J]. Ethics & information technology, 2010, 12 (3): 235-241.

[20] VERUGGIO G, OPERTO F. Roboethics: social and ethical implications of robotics[M]. Berlin: Spinger, 2008.

[21] HANSEN K L, ZENOBIA K E. Civil engineer´s handbook of professional practice[J]. 2011, 75(1):1-2.

[22] KARDOS-KAPONYI E. The charter of fundamental rights of the European Union [J]. Official journal of European Communities, 2000, 23(1/2): 137-170.

[23] PARRISH J L. Papa knows best: principles for the ethical sharing of information on social networking sites [J]. Ethics & information technology, 2010,12(2): 187-193.

[24] White House Office of Science and Technology Policy. Preparing for the future of artificial intelligence [EB/OL]. (2016-10-01) [2018-07-13]. https://obamawhitehouse.archives.gov.

[25] 格德门德尔·阿尔弗雷德松.世界人权宣言[M]. 中国人权研究会,译. 成都: 四川人民出版社,1999.

[26] PAPERNOT N, ABADI M, ÚLFAR ERLINGSSON, et al. Semi-supervised knowledge transfer for deep learning from private training data[EB/OL]. (2016-10-01) [2018-07-13]. https://www.researchgate.net.

[27] HUTSON M. How artificial intelligence could negotiate better deals for humans [EB/OL]. (2017-09-11)[2019-06-30]. https://www.sciencemag.org/news.

[28] MISHRA S, STURM B L, DIXON S. Introduction to local interpretable model-agnostic explanations[EB/OL]. (2016-08-12) [2018-07-13]. https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime.

[29] VOOSEN P. How AI detectives are cracking open the black box of deep learning[EB/OL]. (2017-07-06)[2019-06-30]. https://www.sciencemag.org/news.

[30] DAVID G. Explainable artificial intelligence (XAI)[EB/OL]. (2016-10-16) [2018-07-13]. https://www.darpa.mil/program/explainable-artificial-intelligence.

[31] DOM G, JOLENE C. Google deepmind shows that AI can have "killer instincts"[EB/OL]. (2017-02-13) [2018-07-13]. https://futurism.com/google-deepmind-researchers-show-that-ai-can-have-killer-instincts/.

[32] 高晨阳. 论中国传统哲学整体观[J]. 山东大学学报(哲学社会科学版),1987(1):113-121.

[33] UNESCO. The precautionary principle[EB/OL]. (2005-03-01) [2018-07-13]. http://www.precautionaryprinciple.eu/.

Connotation and Construction Principles of Artificial Intelligence Ethics in the Era of Intelligent Education

DU Jing, HUANG Ronghuai, LI Zhengxuan, ZHOU Wei, TIAN Yang
(Smart Learning Institute, Beijing Normal University, Beijing 100875)

[Abstract] In the era of intelligent education, how man and machine coexist is the key to the ethical construction of artificial intelligence. Firstly, from the perspective of technology paradox, this paper clarifies the ethical challenges and dilemmas existing in the application of artificial intelligence in education, such as technology abuse, data leakage, identity and power boundary of intelligent teaching machines, etc. Secondly, content analysis is used to analyze and extract the ethical elements related to AI ethics by combining the policy documents of many countries and international organizations, and discover the key elements such as value, human interests, security, privacy and responsibility that are repeatedly mentioned in documents of governments, universities and international organizations. Finally, in terms of human-computer co-existence and the application status of AI in the field of education and the key elements of ethics, this paper summarizes the principles to be followed by intelligent education ethics, including the principle of accountability, the principle of privacy, the principle of equality, the principle of transparency, the principle of non-maleficence, the principle of non-independence, the precautionary principle and the principle of stability.

[Keywords] Intelligence Education; Artificial Intelligence; Machine Ethics; International Policy; Principles; APETHICS Model