中国人工智能伦理研究进展

卢艺[1],崔中良[1,2*]

1. 南京信息工程大学马克思主义学院,南京 210044
2. 华东师范大学哲学系,上海 200062

摘要　通过分析目前中国人工智能伦理研究的发展态势,梳理了AI伦理发展历程、伦理风险治理、主体伦理责任及伦理问题成因等4个研究主题;发现研究存在理论与实践脱轨、内容体系性缺乏、研究方向陈旧、研究对象不均衡等不足。对此,紧跟国家人工智能伦理治理动向,着眼AI技术前沿,提出了从重视理论的落地性、鼓励学科研究的体系性、推进伦理研究的创新性、促进研究对象的均衡性等4个方面来构建中国特色人工智能伦理体系,以推动人工智能伦理研究。

关键词　人工智能;伦理风险;伦理治理;体系建构

现代人工智能研究可以追溯至艾伦·图灵发明的图灵机,通用图灵机的问世为1946年第1台计算机的发明打下了基础。1956年夏,达特茅斯暑期学校召开了一次学术会议,有10位人工智能研究者参加,麦卡锡在会上正式提出“人工智能”(artificial intelligent)这一概念[1]。60多年来,信息技术和科学的不断进步推动科技领域出现了许多新变化。虚拟现实、人类增强和人工智能等技术正悄然改变着社会结构和人类的生活方式,同时也引发了不少伦理争议,越来越多的学者对此展开研究,例如徐向东[2]对人类增强技术引起的异议进行了伦理反思,主张“在伦理准则的制约下对基因技术保持接纳的态度;对待新技术的伦理问题应保持理性,要在伦理原则的指导下行动”。随着人工智能与人类生活结合的日益紧密,各领域的专家学者和政府机构的相关人员从不同学术背景出发,展开对人工智能伦理的激烈讨论。

1　人工智能伦理发展历程

第二次世界大战结束后,科学技术的迅猛发展促使人类开始以更理性的态度审视技术,科技伦理在此背景下应运而生。伴随人工智能等新兴技术在各个领域的广泛应用,各界学者也开始在新技术革命背景下思考人工智能伦理的新内涵。中国对人工智能伦理研究的发展历程大致经历了机器人

收稿日期:2021-12-28;修回日期:2022-05-17
基金项目:国家社会科学基金青年项目(21CZX020);国家社会科学基金重大项目(19ZDA033)
作者简介:卢艺,硕士研究生,研究方向为人工智能哲学,电子信箱:20211230021@nuist.edu.cn;崔中良(通信作者),副教授,研究方向为人工智能哲学,电子信箱:cuizhongliang1986@126.com
引用格式:卢艺,崔中良. 中国人工智能伦理研究进展[J]. 科技导报, 2022, 40(18): 69-78; doi: 10.3981/j.issn.1000-7857.2022.18.009

伦理、数据伦理和人工智能伦理3个阶段。

1.1 机器人伦理

机器人伦理在计算机伦理成熟并成为一门独立学科之后逐渐发展起来。关于机器人最早的伦理准则,不得不提及美国作家阿西莫夫,他在小说中首次为机器人设定了3条伦理定律,深刻启发并影响了后人对机器人伦理的研究与讨论。王东浩[3]总结了机器人伦理研究中的问题和困难,并对机器人的道德构建进行整合式分析,扩展了机器人伦理学的原则性内容并提出人机和谐的伦理理念。王绍源[4]指出人工智能技术的发展需要重视机器人伦理的研究及其社会后果,认为自治性智能机器人将在未来引发道德新问题,可能需要其在道德和法律领域承担相应责任。两位学者对机器人伦理学研究现状作了分析总结,并尝试对机器人进行伦理探讨与设计。

1.2 数据伦理

数据伦理是继机器人伦理之后,应用伦理学中的一个新分支。在数字经济全球化背景下,国家越发重视数字经济的发展,制定了有关法规保障数据安全、引导数据合理使用。习近平总书记多次强调要发挥数据资源优势,保障数据安全,实行数字化改革,掌握数字经济发展的主动权。数字经济健康发展,对于打造新发展格局、抢占发展制高点,实现中国经济高质量发展有重要意义[5]。闫宏秀[6]对数字赋能引发的伦理问题进行了反思,主张走出人类中心主义,在人机共融的环境中为数据赋能构建有效的伦理基础设施,补充现有伦理体系,促进伦理新问题的解决,为中国数据伦理发展注入新活力。一些学者认为数据伦理是从数据视角对人类行为的伦理关注[7],他们着眼于大数据时代,在万物数据化的发展趋势下对数据伦理进行思考,严卫等[8]以斯诺登事件为背景,主张多学科讨论、建立多视角解决体系,以开放接纳的心态有序推进大数据伦理问题的解决。而蓝江[9]则主张在数据环世界构建以数据为基础的新伦理法则,寻求多智慧体的共存。

1.3 人工智能伦理

人工智能在交通、医疗等领域的应用过程中出现了许多伦理道德问题,这些问题缺乏理论依据和学科考量,于是学者们开始思考智能时代下伦理学的新内涵。总体来说,学者们较为重视人工智能伦理的建设及相关问题的探讨,例如陈小平[10]提出了人工智能伦理的建设目标、安全底线、功能评价原则、治理责任的落实进路等6个议题,强调在增进人类福祉原则的导向下利用好人工智能技术,完善新时代人工智能伦理体系;王天恩[11]提出智能伦理概念并论述其治理进路,以期更为深入地理解碳基智能体和硅基智能体之间的关系,推动智能社会发展;还有学者基于机器伦理学视角探讨其在应然与实然层面的问题,主张构建“有限人工道德体”(BAMAs)以充分利用深度学习等人工智能技术,同时借鉴人类道德经验提高人工道德体的可实现性和安全性[12]。

由此看来,当前中国人工智能伦理的研究重点有了新变化,逐渐由以人类为基本研究对象向多个研究对象转变。虽然中国关于不同技术维度下对人工智能伦理的探讨开始较早,但是从论文数量和质量上看关注度稍显不足,人工智能伦理研究还未形成完善的结构体系。如今人工智能技术发展势头正盛,世界各国已经开始采取行动并出台政策鼓励和促进人工智能的研发与应用,呼吁AI技术的长远发展应与伦理学相伴而行。与此同时,人工智能伦理发展离不开对人工智能社会影响的研究,尤其2016年AlphaGo完胜韩国棋手李世石等相关新闻报道更渲染了对人工智能技术的伦理担忧,据此引出了人工智能的伦理风险治理问题。

2 人工智能伦理风险治理

伦理风险作为一种新型的社会风险,是指在人与人、人与社会、人与自然、人与自身的伦理关系方面由于正面或负面影响可能产生的不确定事件或条件,尤指其产生的不确定的伦理负效应,诸如伦理关系失调、社会失序、机制失控、人类行为失范、心理失衡等[13]。人工智能伦理风险是指由于人工智能技术应用效果的难预测性和不确定性,以及技术使用者的使用不当等因素而引起的失业、人的异

化、贫富差距加剧等已知或未知的伦理负面影响。在人工智能应用范围扩大、影响深入的情况下,让人工智能在可控范围内为人类带来便利、解放人类劳动,有效控制伦理风险就成为学界讨论的重要主题。中国学者对人工智能伦理风险的治理研究,根据不同的视角大致包括原则规范、理论导向和伦理理念等3个角度。

2.1　原则规范嵌入

原则规范包括原则制定与法律制度的确立。面对当前人工智能伦理风险的加剧,许多学者从伦理问题入手开展对策研究,寻求防控治理人工智能伦理风险的原则规范。其中多数以人类行为为对象,例如在教育和医疗领域内,张寒等[14]从人工智能医疗应用的歧视问题出发,认为人们的态度与行为会诱发伦理风险,相关行动者应确定行为限度、完善伦理规则以严格把控人工智能技术运行的全过程。李晓岩等[15]针对教育管理者、教师及学生三方主体提出人工智能教学应遵循的伦理原则,主张发挥教育主体的能动性,推动人工智能教育的健康发展。李俊平[16]则以更全面的视角论述了一系列伦理新问题,主张以法律与伦理规范相结合的形式防控伦理风险,促进人工智能伦理向善发展,为其他学者的伦理风险研究提供启发和思路。还有少数学者以人工智能产品为切入点并以其发展过程为参考,提出完善法规、制定安全标准、加强国际合作等要求,以此防范伦理风险[17]。对于人工智能应用引起的责任主体模糊、道德主体不确定、人的异化加剧等问题,吴戈[18]从人工智能伦理原则、伦理道德制度和法治体系的顶层设计角度,倡导人与人工智能和谐相处,对人工智能进行道德引导与价值规制,以更好地服务人类。

2.2　理论导向治理

理论导向是指依据某一理论探究人工智能的伦理治理对策。例如杜严勇[19]依据自反性治理理论,从多个层面解释了伦理治理的内涵,介绍了自反性治理的关键问题并论述了相应治理方法,倡导治理者打开认知闭合、强化能力获得以及合作、开放地进行伦理治理,对于科技治理的稳步发展有重要的理论意义。针对人工智能引发的道德失范现象,徐玖玖[20]借助原则导向治理理论寻求调和人工智能技术性与道德性之间冲突的法治策略,以科学-法治-人文为治理思路灵活包容地应对道德伦理风险方案,以实现人工智能的道德性。由于人工智能技术的社会风险在早期具有难预测性,因此,在应用过程中则出现难以控制负面后果的科林格里奇困境,文成伟等[21]以此为切入点探讨了如何运用现代伦理手段消解困境,对于困境引发的社会风险引入预知性技术伦理,着重在技术研发阶段解决问题。

2.3　伦理理念指引

有学者从理念上研究人工智能伦理问题的治理方法。成海鹰[22]从伦理学角度论述了人工智能给人类带来的新问题,以“利益、责任和爱”为中心理念阐明人类应如何在技术时代下保持“在一起”的凝聚力和对人工智能的控制权。还有学者从生命本质入手,对于智能社会过度数据化而引发的生命平等性问题,认为需要重新思考生命本质以及死亡,提出“敬畏生命”“天人合一”等伦理理念,主张“善”用技术以造福人类[23]。可以看到,学界主要从现有伦理问题出发提出相应治理理念,侧重点各有不同,但是对不同领域的伦理理念进行整合性研究的文章较少。

除了以上3个角度外,还有学者从内部进路入手,对伦理治理方法展开探讨,例如王钰等[24]根据技术的发展阶段,对人工智能进行道德化设计,强调作为伦理性道德能动体的技术专家、评估委员会、政府部门以及产品使用者应主动承担责任,发挥人工智能的伦理潜能。值得一提的是,吴红和杜严勇[25]提到了伦理原则如何与实践接轨、落到实地,针对人工智能的社会伦理风险,总结当前主要的伦理治理原则及其疏漏之处,阐明将人工智能伦理治理原则落到实处的行动方向,对多个行为体提出伦理要求,重点关注了人工智能的伦理原则“落地”问题。这一定程度表明中国的人工智能伦理研究正越来越关注伦理实践。

总的来看,学者对各自领域(如教育学、医学、法学等)内人工智能所引发的伦理问题作了较为全面的分析,提出针对性的风险治理方案。由于学者

们集中于人工智能风险对人类自身利益影响的考察,因而普遍主张以伦理原则和规范来约束人工智能,提倡应理性地看待并负责任地使用AI。这些研究往往从具体理论或特定角度探索实现人工智能道德性的治理进路,却很少进一步研究理论的实际性。此外,对于伦理治理的空间与性别异质性问题还缺少专门研究,例如不同性别和种族文化对伦理治理的态度与表现是否存在不同。对于治理对策,学者大多通过风险的表现分析、制定和总结伦理原则,往往停留在理论层面。实际上,伦理原则只是一种抽象规则,还无法被人们具体地应用到实践中去。迈凯伦(Bruce M. McLaren)[26]在对伦理学原则的研究中指出:专家们所提出的原则性建议虽然可能是令人信服且有道理的,但却未必是权威的,因为他们提出的建议措施往往是抽象的,对于这些原则的具体内涵,还缺少详细的内容性解释。

3 人工智能伦理主体责任

人工智能伦理的风险治理,离不开对伦理责任主体问题的探讨。人工智能伦理主体,是指人工智能技术应用及其产品的设计、开发、使用和监管所涉及的各类责任主体。人工智能作为人造物,是人类社会发展的工具,它所带来的一系列不良影响与后果必然需要由创制者负责。而且,虽说人工智能正在朝向强人工智能的方向发展,但依旧处在弱人工智能阶段,自我意识暂未形成、无法独立承担行为后果而且不具备责任主体的承担能力,“奇点”何时到来还难以预知,因此,人类在每一阶段都需要对人工智能行为引起的伦理问题负责,以尽量避免或减轻人工智能技术引发的伦理风险及后果。责任,在《现代汉语词典》中的释义为分内应做的事以及没做好分内事而应承担的过失,显然,人工智能责任的内涵应该取后者。主体伦理责任是指由于人工智能技术的广泛应用而带来人的异化、不平等加剧、“无用阶级”的产生等一系列伦理问题及其负面影响,不同行为主体对此需要承担责任。经过文献的阅读、梳理和分类,发现中国对于人工智能主体伦理责任的研究主要有4个角度:法律、哲学、教育和医疗。

3.1 法律责任角度

随着人工智能产品在人类生活中的广泛运用,人工智能技术引发的法律责任问题逐渐凸显。余婷[27]从法律层面阐述了包括工程师和制造商在内的两类主体为人工智能引发的“恶果”所需承担的责任,她认为人工智能体尚不具备自主能力,带有利益需求的设计者或厂商在很大程度上影响或决定着人工智能体的行为,因此其有义务对人工智能产品的行为后果负责。党家玉[28]探讨了人工智能技术引发侵权问题后对于人工智能的法律责任的具体归属问题,主张明确具体责任承担者,平衡技术使用、生产和受害主体的责任,合理控制风险的同时保持人工智能发展的活力,同时还建议应借鉴国外的经验等。

3.2 哲学责任角度

中国有关人工智能问题的哲学探讨自20世纪70年代便已出现,但受到当时人工智能技术发展水平的制约,学界主要是从人工智能对哲学发展意义方面的理论探讨。由于人工智能技术有了更广泛的应用和伦理影响,学者们也随之挖掘出新的研究方向与思路。例如王绍源等[29]从“物伦理学”的哲学角度着手,主张在机器人不断智能化的条件下通过伦理设计来提升机器人的道德,人为地使其承担一定的伦理责任,与此同时不能忽视每个设计、开发或使用人工智能技术的人类主体的责任。杜严勇[30]从道德哲学的新伦理学角度出发,主张科研人员应发挥道德想象力,积极承担前瞻性的伦理责任,有道德地设计人工智能,从而使人工智能更好地服务人类。

3.3 教育责任角度

如今“教育+人工智能”已经成为中国现代教育发展的新形态,国家也出台了一系列人工智能发展的政策和战略,尤其重视人工智能教育的推广。学者专门对人工智能教育的伦理问题进行了研究。王前等[31]探究人工智能应用中不同使用者以及人工智能体应负有的责任,阐述了伦理责任存在的问题及可能的伦理风险,主张人类应积极承担自身的伦理责任,重视人工智能使用中的伦理风险,促进

人机关系和谐发展。苏明等[32]基于马克思技术观对人工智能教育进行伦理审视,指出由于智能产品自身的不可控性及其有限的自主性等问题的存在,人工智能教育伦理责任归属的界定面临很多困难,并且强调校方要引导人工智能教育向善发展。

3.4　医疗责任角度

自古以来寻求健康和长寿是人类不懈追求的目标,人工智能为医疗卫生领域的发展提供了技术与理论基础,在诊断病情、评估疾病风险、药物研发等方面发挥巨大作用。然而,它也导致了不少伦理问题,引起医学、哲学等领域的学者对责任问题的重视。学者们高度重视医疗中患者的隐私安全,强调生产者和使用者在智能医疗过程中所应承担的相应伦理责任。陈安天等[33]指出医学诊疗应用中存在患者不信任问题,阐明了医生、医疗机构和人工智能厂家的责任内容,并要求前瞻性地开展人工智能医疗行为的伦理反思、完善有关规则和增强伦理意识。除了智能医疗技术在伦理责任上的强化问题,张荣等[34]还论述了人工智能医疗的数据安全和主体争议问题,主张将责任承担的重点放在设计者、用户与公众上,更全面地探讨了责任主体的范围,要求主体提高伦理责任意识,并符合道德规范地使用智能产品。

总的来看,关于人工智能主体伦理责任的研究还有很大完善空间,主要局限于人工智能在某一应用领域的主体责任研究。学者们普遍认同科研人员应具有人工智能研究的伦理责任,注重预防性主体伦理责任的探讨,但主要还是从现存伦理现象中确定主体责任且学科指向性明显,研究角度仍显单一,缺少对人工智能主体伦理的整体性、全局性和体系性研究。如今人工智能的发展步伐越来越快,各国对于伦理责任问题也予以极大重视,例如欧盟在2011年就明确提出“负责任创新”的科技发展理念。有学者研究和总结了欧盟的这项政策,从中国国情和制度出发探讨如何借鉴其政策的运作方式,建立中国特色的负责任科技创新机制[35]。对于人工智能伦理责任的研究,除了关注国外动态,研究外国先进成果与经验并加以借鉴之外,更应从本国实际情况出发去研究各类主体的伦理责任,努力构建具有中国特色的伦理责任研究体系,丰富相关理论成果,引导人工智能伦理的良性发展。

4　人工智能伦理问题成因

人工智能的伦理风险主要探讨人工智能在微观或宏观层面所带来的社会伦理风险及其产生原因,很多学者对此进行了多角度分析。对于人工智能的伦理问题成因分析,在中国还属于比较新的研究方向,主要包括技术内部因素、外界环境因素以及内外整合因素3个方面。

4.1　技术内部因素

技术内部因素是指从人工智能技术自身的算法、功能特性和用途等方面去考察人工智能所引起的各种伦理影响及其后果的成因。学者们从技术特质以及内部运作上对不同方面的伦理问题产生的原因进行了分析,例如张之沧[36]探讨了人工智能技术具体是如何影响和改造人类的亲情、婚姻和性爱关系的,他指出智能伴侣等新的情爱形式正在通过人工智能技术的完美人格、开放平等性等特质瓦解传统的家庭组成、两性关系、血缘亲情。人工智能技术的伦理缺失正在给个人、社会发展甚至人类生存带来不可忽视的伦理危险。王东等[37]指出当前AI技术的不成熟性容易导致失控,进而造成社会性的伦理风险,主张以人类共同利益为最高规范,改善技术算法设计,以坚持人的主导地位为前提重建人机伦理关系。唐代兴[38]则从技术的本质和功能入手,指出在人机紧密结合的趋势下,社会结构的公正性正在被弱化,因此主张有限制地开发技术,否则人类最终可能陷入自我毁灭的危险中。

4.2　外界环境因素

人工智能伦理问题的外界环境因素是指从技术之外的法规制度、社会观念、主体素养等外部环境探讨人工智能伦理问题的成因。外界环境论者认同伦理制度和道德规范的落后加剧了人工智能的伦理风险,给人类生活的各个方面都带来了伦理冲击。其中,肖杰[39]指出技术人员与使用者的价值观念不统一、主体道德素养不强以及监管不足使得人工智能的伦理风险增加,伦理困境得不到解决。

人类对技术的过度依赖以及法律的不完善也一定程度上加剧了隐私泄露、婚姻观念变革等伦理问题[40]。王武斌[41]指出规范制度更新的滞后容易引发技术与人类伦理道德发展的不平衡,且大众对人工智能尚存在认知的不全面性,这也会导致人工智能伦理问题的发生。

4.3 内外整合因素

有学者将技术自身与外部各因素结合起来,以更为全面的视角探讨人工智能伦理问题的成因。谷雨[42]从技术发展的局限、伦理规则的缺失和监管体系的不完善等3个方面综合分析了人工智能带来的各类伦理问题的原因,希望从这些原因出发寻求合理的解决方案,以此规避伦理风险带来的损害,促进人工智能的健康发展。邓若玉[43]对于人工智能技术发展导致的伦理维度偏差问题,从技术的算法逻辑局限、技术制约、伦理道德约束力等内外多个方面阐述了人工智能伦理问题的成因,强调中国进行伦理规范方面研究的紧迫性和重要性。两位学者对于各自探讨的伦理问题,都综合考虑了外部力量和技术内部的因素,将技术缺陷和外部约束力作为共同成因,强调对症下药以加快伦理困境的解决,完善人工智能伦理的现行发展体系。

由此可见,在技术内部方面,中国的学者从家庭构成、社会公正以及人类生存等现有伦理问题切入,以技术的功能及其特性作为伦理问题产生的重要原因,并具体分析人工智能如何导致这些矛盾的产生;对于外部因素,伦理规范与法制的不完善成为普遍共识,人们的认知水平和伦理观念是被主要讨论的问题之一;对于内外整合因素,学者们往往将技术内部和外界制约作用结合来分析,研究角度和范围更为全面。概括说来,学者们主要针对特定伦理问题进行成因分析,分析角度相对分散。不少学者会把成因与应对策略联系起来共同讨论与反思,其实我们更需要对伦理风险问题进行单独地、系统全面地探究。总的来说,风险成因研究在中国仍处于初期阶段,对人工智能伦理风险成因的溯源和探讨上存在一定的重复性,对于人工智能伦理问题成因研究的广度和深度还有待进一步提高。

5 人工智能伦理研究展望

通过对当前的研究进行梳理和总结得出,关于人工智能伦理的论文数量逐渐增多且出现许多新角度,例如从中国传统理念入手研究人工智能伦理现象及其伦理设计等;主要以人类福祉为根本出发点进行伦理研究,且集中在伦理风险治理和理论设计方面;少数学者针对人与人工智能如何相处的问题进行反思,研究人机和谐共处的对策。例如王萍萍[44]以《老子》的"善"的思想为依据,以"善"为理论核心反思人类对人工智能机器人的伦理关怀。不同领域的专家学者的研究成果推动了中国人工智能的伦理研究进程,对当代人工智能的伦理分析大有裨益。但不难发现,国内研究依旧存在一系列问题,还可以在以下方面进一步推进。

5.1 理论的落地性

不论是人工智能风险治理还是主体责任研究,虽然学者们从社会现象以及伦理问题出发,有针对性地探讨了如何从原则、理念和理论导向上有效指引人工智能的正向发展、规避伦理风险,然而多数学者忽视了理论成果如何"落地"的问题。无论是治理对策还是主体责任的确定,不仅需要抽象理论的"对症下药",更要关注理论"实际化"的途径。目前,不少原则比较模糊和抽象,理论应用的难度较大,再完善的理论若无法转化为具体的行动只能是空中楼阁,因此需要研究者能够全过程地参与人工智能研究。在深度神经网络技术蓬勃发展的背景下,深度学习算法逐渐使人工智能脱离人为因素影响,深度挖掘数据特征[45]。可以利用基于深度神经网络的学习算法,在给定标注过的数据集或者伦理案例后,使人工智能对此进行反复学习来获取伦理规则,并在此过程中不断自我识别与矫正,同时设计者要检查和修正人工智能模型成果,尽可能避免数据偏见。另外,理论转化为实践还需要考虑不同地区的文化习惯和取向等,使之能适用于不同的文化环境。设计者可以据此为人工智能制定多样化的行为规范,灵活地进行干预和规则的调整,因地制宜地对人工智能开展伦理教育和纠正。通过自

上而下与自下而上相结合的学习方式,强化伦理理论的灵活性和适应性,进而提高成果的实际应用效果。因此,建议从深度学习这一技术层面着手对伦理理论进行落地性研究,既保证理论不是无法实施的“空谈”,也保证它们可以被“区域化”接受。

5.2　内容的体系性

随着人工智能社会影响力的加深,人工智能伦理问题在逐渐渗透到不同的社会领域,更加证明了构建人工智能伦理体系的紧迫性。而目前学者们更多关注某一具体领域或学科内的问题,在伦理责任主体界定上,普遍从现存伦理问题中寻找答案且有很强的学科指向性,仍缺少整体性的、全局性的体系设计。又如在伦理风险治理层面,学界因专业差异而具有不同的判断角度,很容易导致研究出现行业沟壑或断层,这将会阻碍中国人工智能伦理研究的持续推进。2020年,国家自然科学基金委员会为进一步推动学科交叉融合、鼓励科技发展,专门成立交叉科学部,包括综合与战略规划处、交叉学科一处、交叉学科二处、交叉学科三处和交叉学科四处。其中,交叉学科二处专门研究大数据、人工智能等交叉学科,目的是满足加快国家经济转型的需要,破解经济高质量发展过程中的重大技术难题,强化国家战略科技力量。人工智能作为跨计算机科学、信息科学、语言学和哲学等多学科的交叉学科,本身就带有跨学科性,因此,建议首先从教育、医疗、生物、心理、环境等多个学科与人工智能伦理研究进行有机融合,在开发新学科领域的同时继续对原有学科的伦理进行体系性研究;其次,构建多学科、多领域的规范化学术分享机制,定期组织各领域专家学者进行人工智能伦理的学术交流,鼓励学者以交叉学科的思维和交流互鉴的态度展开人工智能伦理的全方位研究,推动研究的全面性和体系性,依靠外在机制与内部交流支撑人工智能伦理的平衡发展;最后,还要借鉴各学科领域的最新研究成果和国外研究经验,基于中国人工智能实际发展情况,针对性地发现和解决问题,重视人工智能伦理中国特色发展体系的塑造。总之,需要各领域、各学科的专家学者共同交流彼此的研究成果,及时更新各领域研究进展,扩展伦理研究机制,构建完整的人工智能伦理研究和应用体系。

5.3　方向的创新性

当前人工智能伦理的研究方向存在重叠,如伦理问题成因的分析角度相对陈旧、应对法规和原则在内容上缺少创新。在风险治理层面上,主要是从人工智能到人类的单向度对策研究,建议从人工智能主体入手采取双向度的治理研究,以更开放的态度对待人工智能的身份,走出人类中心观或许有助于人工智能伦理研究寻找新思路。另外,在世界文化相互碰撞、性别问题愈发敏感的环境下,将种族文化与性别因素纳入人工智能伦理研究具有很强的现实意义,国外已有学者展开相关探讨,而国内研究尚处初始阶段。由于不同种族的文化及其社会发展状况(例如宗教、社会整体技术水平等)存在差异,大众对人工智能的认知及社会接受度、人工智能产品及其伦理设计、人工智能伦理的规则制定及治理都会受其影响。如日本受到儒家思想、泛灵论的影响,其设计出的机器人倾向于展示具有亲和力的拟人化特点[46]。在人工智能设计和伦理规范制定领域,性别差异问题非常明显。海根多夫(Thilo Hagendorff)[47]提出,相较于男性思维,女性在解释伦理框架时更多以情感为导向而非依赖理性与逻辑,目前的人工智能伦理准确反映出男性思维主导的态势,人工智能研发者的性别比例也处于男多女少的失衡状态。因此,对于人工智能伦理的未来建设,应在尊重和理解不同种族文化背景的前提下展开人工智能设计,除主流西方文化以外还要关注非裔等影响力相对较小的种族文化,以世界性和包容性眼光构建多样化的人工智能伦理,而非一味地寻求普适性治理方案。同时,重视女性对伦理规范设计的思想嵌入,保障和提升女性在人工智能行业中的话语权。此外,认知科学虽与人工智能研究高度相关,但与认知科学相关的人类学、语言学、教育学等学科对于人工智能伦理的参与度并不高,建议从这几类学科入手推进研究的广度与深度。总之,要及时把握人工智能伦理发展的国内外研究进展,以建立中国特色人工智能研究体系为目的,

挖掘新方向,拓展中国科技伦理的发展空间。在泛智能化社会的背景下,人机关系愈加复杂与紧密,甚至出现此消彼长的态势,可以尝试以更为开放包容的眼光探讨人工智能与人的道德、责任分配以及道德设计等问题,更要平衡好人工智能与人的关系,推动人工智能的平稳向善发展。

5.4 对象的均衡性

面对每一个新事物的出现和发展,趋利性促使人们首先从自身利益出发来分析事物。在人工智能伦理研究中,学者也往往以人类得失为切入点探讨伦理问题成因和对策。当前人机交互融合的趋势愈加明显,机器人的种类和功能日益增多,人类与人工智能的相处模式、人工智能的身份都需要重新审视。如交互性机器人的出现使得人类现有社会伦理失效,启示我们以机器、人、社会与世界彼此融合的命运共同体的眼光解决伦理新问题[48]。因此,在人工智能伦理设计的过程中,首先,要监督并确保人工智能设计者严格遵守道德规范,使伦理守则真正渗透进设计行为中;其次,通过一系列技术手段尽可能降低深度学习过程中的"算法黑箱"所带来的人工智能伦理行为的不确定性和不可解释性风险;最后,在有效控制人工智能行为和确保其符合人类伦理的基础上,呼吁社会在一定程度上降低对人工智能的"敌意",以对待新主体的态度看待人工智能,合理开发和利用人工智能主体的功能。例如,在人工智能机器人权利方面,杜严勇[49]对于智能机器人的权利与伦理问题,提出赋予其尊严和有限度的权利;在人工智能伦理设计上,陈凡等[50]对伦理设计起源及其进路的困境进行阐释,夏永红[51]关注人工智能成为道德大师的可能性问题,通过分析伦理设计的难题得出人工智能无法在道德上超越人类的结论。随着人工智能自主性的增强,更需要我们与机器和平共处,给予人工智能体更多实质性的尊重。除了要制定伦理规范、加强理念宣传来引导人们合理使用人工智能之外,也应重视人工智能的道德化设计并给予人工智能体更多人性关怀,鼓励中国学者从人工智能技术自身视角出发进行伦理研究。

6 结论

对当前中国人工智能伦理的研究进展进行了综述,从AI伦理发展、风险治理、主体责任和问题成因4个方面对中国人工智能伦理的研究现状进行了归纳总结,并对未来的走向作出展望,希望作为引玉之砖以推动中国人工智能伦理的深入发展。虽然研究已取得一定进展,但其中存在的问题也不容忽视,例如理论的可实践性、方向的创新性、内容的系统性等都需要进一步完善。

为有效应对中国人工智能伦理的发展需求,应加快构建中国特色的人工智能伦理研究框架,大力推动人工智能伦理理论的"落地性"研究,整合不同领域专家学者的研究成果,鼓励更多学科间的信息互通,以此促进人工智能伦理体系的整体性建构。国外有关人工智能伦理的研究自2005起呈骤升之势,美国、加拿大、日本等国家不仅在国家政策层面走在世界前列并且予以雄厚的资金扶持,在人工智能的多个领域内已逐渐达成统一并制定了一系列发展协议、组成了一些联盟,例如美国谷歌、IBM、微软等多家科技公司成立的人工智能联盟。近几年来,中国也大步踏上人工智能领域发展的快车道,政府正加快步伐对数字赋能,对人工智能研究进行发展规划与治理;在当前数字化转型背景下,基于深度神经网络研究的新一代人工智能在数据精度、计算能力等方面获得了巨大提升,医疗问诊、智能驾驶、智能交互、疫情防控等各领域纷纷尝试数字化发展路径,体现其与人工智能相互交融的态势,侧面证明了科技是社会发展的强大动力,建议人工智能伦理研究应依靠理念与技术的双重驱动,依靠算法手段提升伦理治理效果。继十九大报告中"数字中国"概念的提出,国家愈发重视科技伦理的治理问题,习近平主席强调,科技伦理是科技活动必须遵守的价值准则,要坚持增进人类福祉、尊重生命权利、公平公正、合理控制风险、保持公开透明的原则,健全多方参与、协同共治的治理体制机制,塑造科技向善的文化理念和保障机制[52]。因此,要求中国智能化、数字化发展应顺应国家科技

伦理发展的现实需要,以前瞻性眼光洞察人工智能的发展动态,将国家政策、社会伦理问题纳入人工智能研究中,立足当前并着眼未来,引导人工智能避恶向善。

参考文献(References)

[1] 尼克. 人工智能简史[M]. 北京: 人民邮电出版社, 2017.

[2] 徐向东. 人类增强技术的伦理审视[J]. 哲学分析, 2019, 10(5): 4-29.

[3] 王东浩. 机器人伦理问题研究[D]. 天津: 南开大学, 2014.

[4] 王绍源. 应用伦理学的新兴领域: 国外机器人伦理学研究述评[J]. 自然辩证法通讯, 2016, 38(4): 147-151.

[5] 新华社. 习近平在中共中央政治局第三十四次集体学习时强调 把握数字经济发展趋势和规律 推动我国数字经济健康发展[EB/OL]. (2021-10-19) [2022-05-09]. http://www.news.cn/2021-10/19/c_1127973979.htm.

[6] 闫宏秀. 数据赋能的伦理基质[J]. 社会科学, 2022(1): 136-142.

[7] 颜世健. 数据伦理视角下的数据隐私与数据管理[J]. 新闻爱好者, 2019(8): 36-38.

[8] 严卫, 钱振江, 周立凡, 等. 人工智能大数据伦理问题的研究[J]. 科技风, 2019(28): 105-106.

[9] 蓝江. 从碳基伦理到硅基伦理——人工智能时代的伦理学浅论[J]. 道德与文明, 2020(9): 36-44.

[10] 陈小平. 人工智能伦理建设的目标、任务与路径:六个议题及其依据[J]. 哲学研究, 2020(9): 79-87.

[11] 王天恩. 智能伦理:人工智能时代的伦理新视野[J]. 阅江学刊, 2021, 13(2): 15-24.

[12] 阮凯. 机器伦理学的当代争议及其解决方案[J]. 自然辩证法研究, 2021, 37(11): 42-48.

[13] 陈爱华. 高技术的伦理风险及其应对[J]. 伦理学研究, 2006(4): 95-99.

[14] 张寒, 王锦刚. 人工智能改变医学的理性权衡——基于MIT/MGH乳腺癌筛查人工智能系统的考察[J]. 自然辩证法研究, 2021, 37(1): 40-46.

[15] 李晓岩, 张家年, 王丹. 人工智能教育应用伦理研究论纲[J]. 开放教育研究, 2021, 27(3): 29-36.

[16] 李俊平. 人工智能技术的伦理问题及其对策研究[D]. 武汉: 武汉理工大学, 2013.

[17] 孙保学. 人工智能的伦理风险及其治理[J]. 团结, 2017(6): 33-36.

[18] 吴戈. 人工智能发展带来的问题及其伦理思考[J]. 中州学刊, 2021(3): 93-95.

[19] 杜严勇. 论人工智能的自反性伦理治理[J]. 新疆师范大学学报: 哲学社会科学版, 2018, 39(2): 111-119.

[20] 徐玖玖. 人工智能的道德性何以实现——基于原则导向治理的法治进路[J]. 现代法学, 2021, 43(3): 24-40.

[21] 文成伟, 汪姿君. 预知性技术伦理消解人工智能科林格里奇困境的路径分析[J]. 自然辩证法通讯, 2021, 43(4): 9-15.

[22] 成海鹰. 人工智能时代的"在一起"[J]. 中国人民大学学报, 2021, 35(1): 57-65.

[23] 朱清华. 智能时代人类生命本质的变异及其价值影响[J]. 自然辩证法研究, 2021, 37(2): 124-128.

[24] 王钰, 程海东. 人工智能技术伦理治理内在路径解析[J]. 自然辩证法通讯, 2019, 41(8): 87-93.

[25] 吴红, 杜严勇. 人工智能伦理治理: 从原则到行动[J]. 自然辩证法研究, 2021, 37(4): 49-54.

[26] McLaren B M. Extensionally defining principles and cases in ethics: An AI model[J]. Artificial Intelligence, 2003, 150(1-2): 145-181.

[27] 余婷. 人工智能的伦理问题及对策研究[D]. 武汉: 武汉理工大学, 2016.

[28] 党家玉. 人工智能的伦理与法律风险问题研究[J]. 信息安全研究, 2017, 3(12): 1080-1090.

[29] 王绍源, 赵君. "物伦理学"视阈下机器人的伦理设计——兼论机器人伦理学的勃兴[J]. 道德与文明, 2013(3): 133-138.

[30] 杜严勇. 论人工智能研究中的前瞻性道德责任[J]. 上海师范大学学报: 哲学社会科学版, 2018, 47(4): 43-49.

[31] 王前, 曹昕怡. 人工智能应用中的五种隐性伦理责任[J]. 自然辩证法研究, 2021, 37(7): 39-45.

[32] 苏明, 陈·巴特尔. 人工智能教育伦理的多维审视——基于马克思技术批判和人的全面发展理论[J]. 西南民族大学学报: 人文社科版, 2019, 40(11): 223-228.

[33] 陈安天, 张新庆. 医学人工智能辅助诊疗引发的伦理责任问题探讨[J]. 中国医学伦理学, 2020, 33(7): 803-808.

[34] 张荣, 徐飞. 人工智能医学伦理问题及对策研究[J]. 医学与哲学, 2020, 41(13): 14-19.

[35] 薛桂波, 赵一秀. 基于"负责任创新"的欧盟科技政策转型及启示[J]. 中国科技论坛, 2017(4): 172-177.

[36] 张之沧. 人工智能对家庭伦理的冲击与解构[J]. 国外社会科学前沿, 2021(1): 3-14.

[37] 王东, 王振. 人工智能伦理风险的镜像、透视及其规避[J]. 伦理学研究, 2021(1): 109-115.

[38] 唐代兴. 人工智能发展带动的社会公正危机[J]. 人文杂志, 2020(8): 19-28.

[39] 肖杰. 人工智能技术发展的伦理问题研究[D]. 成都: 成都理工大学, 2020.
[40] 李赫. 人工智能对伦理道德的影响研究[D]. 成都: 成都理工大学, 2019.
[41] 王武斌. 人工智能技术发展的伦理问题研究[D]. 成都: 成都理工大学, 2019.
[42] 谷雨. 人工智能发展的伦理问题研究[D]. 重庆: 西南大学, 2018.
[43] 邓若玉. 人工智能发展的科技伦理反思[J]. 广西社会科学, 2020(10): 93-97.
[44] 王萍萍. 人工智能时代机器人的伦理关怀探析——以《老子》"善"论为视角[J]. 自然辩证法研究, 2021, 37(5): 54-59.
[45] 林涛, 刘爱连. 基于医学影像的影像组学及深度学习在肝细胞癌中的研究进展[J]. 中国医学影像学杂志, 2022, 30(4): 401-405.
[46] Nomura T. Who and under what context requires "Roboethics"? From cross-cultural perspective on assumptions about robots[J]. International Journal of Humanoid Robotics, 2008, 5(1): 25-46.
[47] Hagendorff T. The ethics of AI ethics: An evaluation of guidelines[J]. Minds and Machines, 2020, 30(1): 99-120.
[48] 崔中良, 王慧莉, 郭聃, 等. 人工智能研究中交互性机器伦理问题的透视及应对[J]. 西安交通大学学报: 社会科学版, 2020, 40(1): 123-132.
[49] 杜严勇. 论机器人权利[J]. 哲学动态, 2015(8): 83-89.
[50] 陈凡, 徐旭. 当代人工智能伦理设计的困境和超越[J]. 华中科技大学学报: 社会科学版, 2020, 34(5): 1-7.
[51] 夏永红. 人工智能可以成为道德大师吗[J]. 道德与文明, 2021(1): 106-117.
[52] 新华社. 习近平主持召开中央全面深化改革委员会第二十三次会议强调 加快建设全国统一大市场提高政府监管效能 深入推进世界一流大学和一流学科建设[EB/OL]. (2021-12-17)[2022-05-14]. http://www.news.cn/politics/leaders/2021-12/17/c_1128174996.htm.

Research progress and trend on the artificial intelligence ethics in China

LU Yi[1], CUI Zhongliang[1,2*]

1. School of Marxism, Nanjing University of Information Science & Technology, Nanjing 210044, China
2. Department of Philosophy, East China Normal University, Shanghai 200062, China

Abstract In order to provide useful guidance for further improvement of AI ethics, this paper analyzes the current development trend of AI ethics researches in China. The relevant literature in China is reviewed, focusing on four research topics, including the development process of AI ethics, the governance of ethical risk, and the ethical responsibility of the main body. The causes of ethical problems are analyzed, and an analysis framework is constructed. Through the review, some limitations in the domestic researches are revealed, such as the disconnect between theory and practice, the lack of systematicness, the obsolescence of research direction, and the imbalance of research objects. In view of the above problems, according to the trend of AI ethical governance in China, and focusing on the forefront of AI technology, some suggestions are made: emphasizing that theories should be linked to practice, encouraging the systematic researches of disciplines, promoting the innovation of ethical research, and promoting the balance of research objects. Following the guidance of these four aspects, this paper proposes to build an AI ethics system with Chinese characteristics and promote the research of AI ethics.

Keywords artificial intelligence; ethical risk; ethical governance; system construction

（责任编辑　王丽娜）