重庆邮电大学学报（社会科学版）
Journal of Chongqing University of Posts and Telecommunications(Social Science Edition)

人工智能伦理规范核心原则的政策文本内容分析

周庆山, 郑霞, 黄国彬

引用本文:
周庆山,郑霞,黄国彬. 人工智能伦理规范核心原则的政策文本内容分析[J]. 重庆邮电大学学报（社会科学版）, 2022, 34(5): 9–20.
ZHOU Qingshan,ZHENG Xia,HUANG Guobin. The Core Framework and Content Analysis of Ethical Principles of Artificial Intelligence[J]. *Journal of Chongqing University of Posts and Telecommunications (Social Science Edition)*, 2022, 34(5): 9–20.

DOI:10.3979/1673-8268.20220316002
引用格式:周庆山,郑　霞,黄国彬.人工智能伦理规范核心原则的政策文本内容分析[J].重庆邮电大学学报(社会科学版),2022(5):9-20.

人工智能伦理规范核心原则的政策文本内容分析*

周庆山[1],郑　霞[2],黄国彬[3]

(1.北京大学 信息管理系,北京 100871;2.中国人民大学 实验室管理与教学条件保障处,
北京 100086;3.北京师范大学 政府管理学院,北京 100875)

摘　要:[目的/意义]人工智能技术悄然拉开了第四次工业革命的大幕,为各行各业的发展带来了革命性变化。但该项技术的法律规制和标准规范建设尚不完善,引发了全球对于人工智能伦理原则与规范体系的探索与思考,各国学(协)会相继发布了人工智能伦理规范或政策文件。本研究全面分析和梳理当前认可度较高的人工智能伦理原则的核心框架,旨在为各国制定和完善人工智能伦理规范提供参考。[方法/过程]运用网络调研法、内容分析法和比较研究法,在综合考虑影响力、国别和可获取性等因素的基础上,选取近年各国学(协)会已发布的有关人工智能伦理原则的权威政策文件,提炼出其中涉及的人工智能伦理原则的核心框架和主要内容。[结果/结论]人工智能领域主要包括:造福原则、责任原则、公平正义原则、安全原则、尊重人的自主性原则、隐私与个人信息保护原则、透明性原则和可靠性原则等 8 项人工智能伦理原则。

关键词:人工智能;伦理原则;规范体系

中图分类号:B82-05　　**文献标识码**:A　　**文章编号**:1673-8268(2022)05-0009-12

一、研究背景

人工智能(Artificial Intelligence,AI)是探索、模拟和研究人类学习、思维和知识存储等过程,进而拓展人类智能的技术、理论和方法的新兴学科,属于计算机科学的分支,融合了信息学、控制学、语言学、心理学等多学科领域[1]。近年来,人工智能正以指数级的速度高速增长,但与之匹配的行业规范和管理制度却未能同步发展。针对人工智能的设计与管理,还存在政策取向不清晰、伦理规制缺失、法律法规不健全等问题,这导致了人工智能应用的不确定性,并对人类伦理道德带来了相应的挑战。作为一个涉及多个学科的交叉性话题,围绕人工智能伦理的研究,目前已吸引了包括

* **收稿日期**:2022-03-16
基金项目:国家社科基金项目:信息弱势群体电子公共服务利用障碍及援助机制研究(20BTQ079)
作者简介:周庆山,教授,博士生导师,主要从事信息政策与信息法、信息传播学等研究;通信作者:郑　霞,硕士,主要从事信息政策与信息法、注册式研究报告等研究,E-mail:zhengxia9665@ruc.edu.cn;黄国彬,教授,硕士生导师,主要从事信息资源管理、信息法学等研究。

计算机科学、生物学、行为科学、心理学、医学和哲学,以及图书情报与档案管理等在内的多个学科的研究人员,围绕人工智能在设计环节、测试环节和应用环节等涉及的伦理道德开展研究,各国开始积极探索并纷纷颁布了人工智能伦理原则的行动准则和政策文件。为此,笔者拟运用网络调研法和内容分析法等,以近年来各国学(协)会发布的有关人工智能伦理原则的政策文档为研究对象,对比分析现有人工智能政策文档中涉及的人工智能伦理原则的核心框架和内容构成。

二、国内外研究现状

笔者于2021年12月以“人工智能”并含“伦理”及其对应的英文“Artificial Intelligence”(或“AI”)并含“ethics”为关键词,在CNKI、Web of Science等数据库中检索相关文献,拟从研究内容和研究方法两个方面对关于人工智能伦理规范内容要素的理论研究进行梳理和剖析。

(一)研究内容

全面分析当前关于人工智能伦理规范的已有研究发现,从研究内容的角度来看,学者已提出了包含以人为本原则、公正原则、公开透明原则、知情同意原则、责任原则及安全原则等在内的人工智能伦理规范基本条款(见表1)。

表1 代表性文献中提及的“人工智能伦理规范的内容要素”

人工智能伦理规范的内容要素	代表性文献
以人为本原则	孙伟平[2]、Swarte等[3]
公正原则	Jobin等[4]、邓国民等[5]、张浩等[6]
公开透明原则	蒋洁[7]、杜静等[8]、郑志峰[9]
知情同意原则	李伦等[10]、杨保军等[11]
责任原则	李娜等[12]、Ho等[13]
安全原则	曹建峰等[14]、张荣等[15]、Keskinbora[16]

不过,多数文献探讨的是将人工智能技术应用到其他领域可能引发的伦理问题及应对策略,如人工智能与教育学、医学、传播学等领域结合时应构建的伦理原则,且其提出的人工智能伦理规范内容要素大多是主观推断,而对于现有人工智能伦理规范政策文档所涉及的内容要素关注度不高,致使研究结论的深度与广度均显不足。

(二)研究方法

从研究方法来看,当前已有学者采用内容分析法,以近年来国际或各国发布的人工智能伦理规范的政策文档为研究对象,从中提炼出人工智能社会伦理规范并加以剖析。为进一步梳理现有研究所选政策文档发布主体的类型,笔者选取了6篇应用归纳法开展人工智能伦理规范内容研究的代表性文献,并按照前述5种人工智能伦理政策文档发布机构的类型对其进行分析和评述,结果如表2所示。

表2 6篇应用归纳法开展人工智能伦理规范内容构成研究的样本数量分布(按政策发布主体分类)

发布主体	文献1[3]	文献2[17]	文献3[18]	文献4[19]	文献5[20]	文献6[21]
联合国/欧盟/联合会	5	0	6	2	1	8
各国政府管理机构	15	4	11	1	0	16
学会/协会/研究机构	24	13	5	9	2	25
研讨会组织者	15	0	6	4	0	17
人工智能业界	19	10	8	6	0	20

分析表2可知,从所选研究样本的发布主体来看,文献1至文献6所选研究样本对于前述5类组织机构均有所涉及。但其共性问题是:未考虑人工智能伦理规范政策文档制定主体的差异,而是直接、强行地将不同制定主体发布的人工智能伦理规范的政策文档加以横向比较和内容提炼,这一做法会导致将不在同一影响力层面的文本政策混为一谈,无法确定不同制定主体因自身肩负的社会责任、机构定位等的不同而对人工智能伦理规范的关注点、基本认知、价值判断、动机立场的区别,最终得到的研究结论缺乏说服力、严谨性和可靠性。

三、研究方法

本研究以近年来各国学(协)会发布的人工智能伦理规范文档为分析对象,创新之处有以下几方面。第一,研究方法更加规范严谨。笔者以实证分析为基础,从各国学(协)会发布的人工智能伦理规范的相关政策文档中,更为客观全面地揭示其核心模块和内容构成。第二,研究对象具有高度可比性。在现有同主题的研究中,多数学者在选择人工智能伦理规范的研究样本时,只考虑内容主题是否与人工智能伦理规范相关,而不考

虑制定主体是否在职能和权限上处于同一层级，从而导致将不同影响效力和作用范围的人工智能伦理规范进行比较，缺乏严谨的可比性。本研究所选择的样本，其制定主体均为学(协)会，属于职能和权限相同的同类机构，从而避免将职能和权限存在不同的各类制定主体所制定的人工智能伦理规范进行比较，更好地确保样本的可比性。

具体来说，首先，笔者梳理了人工智能社会伦理规范相关主题的研究论文，并对其中涉及的人工智能社会伦理规范的基本条款进行了预调研；其次，全面调研各国学(协)会的官网，获得其发布的与人工智能社会伦理规范相关的框架、指南与原则；最后，以 Ethical IA Resource[22]、Future of Life Institute[23]、链接人工智能准则平台[24]等网站为线索，对拟选研究样本进行完善和补充，最终选择了 10 款由各国学(协)会发布的有关人工智能社会伦理规范的政策文件作为分析对象。按发布时间排序，这 10 款人工智能社会伦理规范依次是：《机器人和机器人设备：机器人及机器人系统的道德设计及应用指南》(Robots and Robotic Devices：Guide to the Ethical Design and Application of Robots and Robotic Systems)、《算法透明度和问责制的声明》(Statement on Algorithmic Transparency and Accountability)、《AI 政策原则执行摘要》(AI Policy Principles Executive Summary)、《人工智能和机器学习：政策文件》(Artificial Intelligence and Machine Learning：Policy Paper)、《人工智能和数据分析的伦理原则》(Ethical Principles for Artificial Intelligence and Data Analytics)、《日本人工智能学会伦理准则》(The Japanese Society for Artificial Intelligence Ethical Guidelines)、《以伦理为基准的设计：人工智能及自主系统以人类福祉为先的愿景(第一版)》(Ethically Aligned Design：A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems(First Edition)、《人工智能职业道德和行为准则》(AAAI Code of Professional Ethics and Conduct)、《负责任的人工智能政策框架》(Responsible AI Policy Framework)、《医疗保健中的增强智能》(Augmented Intelligence in Health Care)，详见表 3 所示。在此基础上，运用内容分析法和比较研究法，对上述政策文档涉及的人工智能伦理原则的内容构成进行分析比较，剖析当前具有较大国际影响力的人工智能伦理原则的核心框架和基本内容。

表 3　10 款各国学(协)会发表的人工智能伦理原则的指南政策基本信息

政策编号	制定机构	政策名称	发布时间	有效链接
1	英国标准学会(British Standards Institution，BSI)	*Robots and Robotic Devices：Guide to the Ethical Design and Application of Robots and Robotic Systems*	2016	https://shop.bsigroup.com/ProductDetail?pid=000000000030320089
2	美国计算机协会(Association of Computing Machinery，ACM)	*Statement on Algorithmic Transparency and Accountability*	2017	https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
3	美国信息技术工业协会(Information Technology Industry Council，ITI)	*AI Policy Principles Executive Summary*	2017	https://www.itic.org/dotAsset/50ed66d5-404d-40bb-a8ae-9eeeef55aa76.pdf
4	国际联网协会(Internet Society，ISOC)	*Artificial Intelligence and Machine Learning：Policy Paper*	2017	https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf
5	美国软件与信息工业协会(Software and Information Industry Association，SIIA)	*Ethical Principles for Artificial Intelligence and Data Analytics*	2017	https://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2019-11-23-152702-120

续表

政策编号	制定机构	政策名称	发布时间	有效链接
6	日本人工智能学会（Japanese Society for Artificial Intelligence，JSAI）	*The Japanese Society for Artificial Intelligence Ethical Guidelines*	2017	http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf
7	美国电气电子工程师学会（Institute of Electrical and Electronics Engineers，IEEE）	*Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)*	2019	https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf
8	国际先进人工智能协会（Association for the Advancement of Artificial Intelligence，AAAI）	*AAAI Code of Professional Ethics and Conduct*	2019	https://aaai.org/Conferences/code-of-ethics-and-conduct.php
9	国际技术法协会（International Technology Law Association，ITechLaw）	*Responsible AI Policy Framework*	2019	https://www.itechlaw.org/sites/default/files/ResponsibleAI_Policy-Framework.pdf
10	美国医学会（American Medical Association，AMA）	*Augmented Intelligence in Health Care*	2019	https://www.ama-assn.org/system/files/2019-08/ai-2018-board-policy-summary.pdf

注：为节省篇幅，本文后续部分仅用表3的政策编号代替政策名称。

四、人工智能伦理规范的核心框架与内容剖析

（一）人工智能伦理规范的核心框架

在对本研究所选择的10款有关人工智能伦理原则规范文档进行预调研的基础上，参照与人工智能领域存在交叉的计算机科学、信息技术[25]和医学[26]等领域构建的基本伦理原则，笔者设计了造福原则、责任原则、公平正义原则、安全原则、尊重人的自主性原则、隐私与保护原则、透明性原则和可靠性原则等8项人工智能伦理原则。在此基础上，统计分析了本研究所选择的10款人工智能伦理政策文件中所涉及的伦理原则数量（见表4）。从表4可以看出，当前人工智能伦理政策文件中多强调造福原则、责任原则、公平正义原则、安全原则、尊重人的自主性原则和隐私与保护原则等，部分政策文件还补充提出了透明性原则、可靠性原则和共同治理原则等内容。

表4　10款人工智能伦理政策文件中涉及的伦理原则内容统计

伦理原则	政策编号									
	1	2	3	4	5	6	7	8	9	10
造福原则		√		√	√	√	√	√	√	√
责任原则	√	√	√	√		√	√	√	√	
公平正义原则		√		√		√		√	√	√
安全原则	√	√	√	√	√	√	√	√	√	√
尊重人的自主性原则	√		√	√			√			
隐私与个人信息保护原则	√			√		√		√	√	
透明性原则		√	√	√	√		√		√	√
可靠性原则	√		√	√		√		√		
其他				√		√	√			

注：“√”表示人工智能伦理政策规范中提及了该伦理原则。

（二）人工智能伦理规范的内容剖析

人工智能伦理原则的提出，旨在为人工智能的开发者、使用者和治理者提供应当遵循的理念和规则。为进一步探索人工智能伦理原则的内容

构成,笔者全面研读了本研究所选择的 10 款人工智能伦理原则政策文件,并对其中所涉及的人工智能伦理原则的内容表述进行了对比分析(见表 5)。尽管不同政策文件中,对人工智能伦理原则内容描述的详略程度和划分维度有所区别,但尽可能全面地涵盖人工智能伦理原则的核心架构,并将其深刻嵌入到人工智能产品、技术和服务的研发、应用和治理等各个阶段,是各国学(协)会设计和构建人工智能伦理原则的基本出发点和最终归宿。

表 5　10 款人工智能伦理原则政策文件对人工智能伦理原则的内容表述

伦理原则	政策 1	政策 2	政策 3	政策 4	政策 5
造福原则	—	人工智能监管机构应制定相关工作机制,为受到不利影响的个人或群体提供救济	—	人工智能开发者致力于保障人工智能产品为人类社会提供福利	人工智能开发者有责任利用其所提供的产品和服务,推动社会发展
责任原则	人类应当对人工智能负责	即使无法对算法决策过程做出合理解释,人工智能开发机构仍需对其所使用的算法负责	人工智能研究人员、学科专家和利益相关者应尽可能确保负责任地设计和部署人工智能系统	人工智能开发者负责随时监测人工智能系统的安全性,并明确人工智能系统的责任分配	—
公平正义原则	—	充分了解人工智能系统设计、使用和实现过程中可能存在的算法偏见及潜在威胁	—	人工智能系统不应当使用存在偏见、不准确、不完整或具有误导性的数据作为决策支持	—
安全原则	人工智能系统应安全运行且符合设计目标	系统涉及的模型、算法、数据和决策都应被精准记录,以便在发生未知风险时对系统进行审核	技术人员需确保人工智能系统安全运行,将保障用户和第三方安全使用视作首要任务	将人工智能系统的安全性放在首要和突出地位	人工智能开发者应及时制定相关政策和评估程序,综合考虑不同情况,定期评估拟使用的算法和数据对部分弱势群体造成的潜在和实质性损害
尊重人的自主性原则	人类有权自主选择人工智能的产品或服务	—	人工智能可为人类提供决策支持,但人类仍应把握人工智能的控制权力	用户有权充分了解算法的运作及决策过程,进而判断哪些功能不适合委托给人工智能进行决策	—

续表

伦理原则	政策 1	政策 2	政策 3	政策 4	政策 5
隐私与个人信息保护原则	人工智能系统应充分尊重个人隐私	—	—	人工智能开发者必须安全、合理地使用数据，遵循必要的隐私政策和个人数据保护条款	—
透明性原则	—	人工智能系统和机构应当充分解释算法运行依据的程序和具体决策过程	借助一定的工具提高人工智能系统的可解释性	人工智能所做出的决策应该具有可解释性，特别是当决策引发公共安全或导致歧视行为时	定期为公众提供系统有效性和可靠运作的监测结果，以帮助公众更好地理解和预测人工智能系统的决策及影响
可靠性原则	人工智能系统不发生欺骗性行为	—	确保开发和改进人工智能系统所需数据的鲁棒性、代表性和可用性	在架构和规划人工智能系统功能时，需要谨慎考量系统指令和算法程序	—
其他	—	—	—	促进人工智能利益相关者共同参与到人工智能的治理过程中	—

注：“—”表示无相关内容，下同。

表 5(续)　10 款人工智能伦理原则政策文件对人工智能伦理原则的内容表述

伦理原则	政策 6	政策 7	政策 8	政策 9	政策 10
造福原则	人工智能开发者应当为人类的和平、安全、福利和公共利益做出贡献	人工智能开发者应将增进人类福祉作为发展宗旨	人工智能研究者的主要职责是为人类社会造福	人工智能系统的开发、部署和使用，应确保符合道德标准，增进人类福祉，且遵循其他人工智能政策框架的基本原则	人工智能的发展应当以用户为中心
责任原则	随时监测人工智能系统性能，一旦发现潜在风险应及时告知社会公众	开发和部署人工智能系统之前，需明确责任主体的权利及义务	—	人工智能系统的开发、部署和使用应当遵守人工智能政策框架中的问责原则	—

续表

伦理原则	政策 6	政策 7	政策 8	政策 9	政策 10
公平正义原则	确保人工智能系统无不平等或歧视倾向,并作为一种公共资源为人类提供平等服务	—	人工智能研究者应保障社会公众平等享有人工智能服务或产品	保障人工智能运行结果的非歧视性,同时要采取恰当和有效的措施维护人工智能使用过程的公平性	人工智能系统要规避歧视问题,关注弱势群体需求
安全原则	认识到保护人工智能安全的必要性,随时关注人工智能的安全性、可控性和保密性	防范人工智能运行中可能出现的误用风险	在设计和部署人工智能系统的过程中,首先需要考虑系统的安全性,制定必要的政策措施保障系统资源的安全,防止系统被误用或篡改等	采用合理的设计制度和运行标准,一方面应确保人工智能系统的安全性和可靠性,另一方面要约束开发者和部署者行为规范	提高人工智能系统的安全性和可重现性
尊重人的自主性原则	—	人工智能的设计和运作应尊重、促进和保护国际公认的基本人权	—	—	—
隐私与个人信息保护原则	在研究和开发人工智能时要尊重他人的隐私,并按照相关法律法规适当地处理个人信息	—	人工智能开发者应充分保护隐私安全,明确用户数据收集的范围、形式、权利和责任	确保人工智能系统符合隐私规范	—
透明性原则	—	制定人工智能开发过程透明性的评估标准,为开发者提供决策指导	—	人工智能的使用过程开放透明,且其决策过程和结果是可解释的	确保人工智能系统的透明性
可靠性原则	以科学、严谨、可靠的方式开展人工智能系统的开发和设计	—	确保向使用者提供的人工智能系统性能、服务质量、潜在风险等信息真实可靠,不得提供虚假或误导性的声明,不得伪造数据资料,实施行贿受贿及其他恶意行为等	—	—
其他	人工智能开发者应当与社会公众建立长效沟通机制	人工智能开发者需具备该技术安全、有效操作所需要的知识和技能	开发人员制定相应标准作为有效性的衡量指标	—	—

1.造福原则

造福(well-being & benefit)原则指的是,人工智能的发展成果应当被全人类所共享,改善人类赖以生存的社会和环境,推动人类和社会繁荣进步,增进人民和社会福祉[6]。该伦理原则指明了人工智能发展的首要目标和最终归宿,为人工智能研发人员在构想、设计和优化人工智能功能模块和系统程序时提供了基本遵循。经分析,本研究所选择的10个分析样本中,有8款人工智能伦理原则的指南规范均明确提出,人工智能的研发需遵循造福人类的基本原则。如政策4指出,人工智能开发者应加强对话和协商,制定人工智能发展战略,保障人工智能为社会公众带来福祉,借助人工智能解决教育改革、收入分配及社会服务等问题,并尽可能规避其不利影响;政策6和政策9均提出,人工智能创造的经济价值应当造福全人类,并为全人类的和平、安全、美好和共同利益做出贡献。此外,政策8和政策10提出,设计人工智能时应当以用户为中心,并最大限度地保护基本人权和人类自主权。在此过程中,人工智能开发者的根本目标是将人工智能给人类带来的健康、安全和隐私等方面的威胁最小化;同时,人工智能的发展还要充分保护人类环境,促进全球自然环境的可持续发展。

2.责任原则

责任(responsibility & accountability)原则指的是,在研发人工智能时应明确人工智能系统、产品、服务或功能的责任主体、责任分配、关联责任,以及当发生不利后果时各责任主体所需承担的后果和义务。该伦理原则规定了开发者在人工智能设计和研发阶段务必遵循的法律法规和社会责任,进一步明确了人工智能发展过程中的责任红线。本研究所选择的10个分析样本中,有7款人工智能伦理原则的指南规范均提到了人工智能责任伦理原则。其中,政策1、政策2和政策6均指出,人类应当对人工智能负责,并为每个人工智能产品或服务设立特定责任人。政策4和政策7对这一原则做出了更为详尽的阐述,认为当下应确保人工智能相关法律的明确性。政府需及时制定适用于算法决策和使用的规制框架,保障法律运作的确定性,将用户权益作为第一要义,并要求人工智能的设计者、开发者和使用者加以遵循;此外,当前需要进一步明确人工智能系统的责任分配,明确规定如若发生安全问题,将由哪些主体承担责任,以及后续的救济或赔偿措施。政策9建议在人工智能系统的开发、部署和使用过程中设立问责机制,即无论在什么情况下,人类都应当对人工智能的行为及后果负责。

3.公平正义原则

公平正义(fairness & justice)原则指的是,以平等、公正的方式设计、分配人工智能的产品或服务,进而形成全社会平等参与、公平使用人工智能技术成果的良好风范。该原则规定,人工智能的训练数据和决策算法不应嵌入开发人员观念中固有的偏见或歧视想法,而应当确保人工智能的产品和服务被所有人平等享有,鼓励提升社会文化的多样性与包容性。本研究所选择的10个分析样本中,有6款人工智能伦理指南规范提到了公平正义原则。具体来说,政策2指出,人工智能的开发者、设计者、架构师及其他利益相关者应充分了解系统设计、实现和使用过程中可能存在的算法偏见,以及这些偏见给人类社会带来的潜在威胁。政策6和政策9均指出,人工智能的使用可能会给社会带来各种性别、年龄、种族等的歧视问题,因此技术开发者在设计人工智能时,应当尽可能避免歧视和偏见,确保该技术以公平的方式被人类使用。政策8阐述了公平正义原则的含义及解决措施,认为这一原则的含义包括平等、宽容、尊重他人、公平正义等。人工智能研究者应保障社会公众平等享有人工智能服务或产品,不因年龄、肤色、身体状况、种族、家庭状况、性别、国别、宗教信仰等差异而受到歧视。鉴于此,人工智能的信息技术应具有包容性和可获取性,尽可能避免对公民权利带来的歧视或不公平行为。

4.安全原则

安全(security)原则指的是,在任何情况下都应当阻止人工智能无意识行为伤害人类,或侵犯人类权益[5]。该原则强调人类应采取审慎态度对待人工智能的产品和服务,避免使自身承担任何不恰当、易受伤的风险[27],确保人工智能在使用过程中安全运作和稳定可靠。本研究所选择的10款人工智能伦理原则指南规范均提到了安全原

则。其中,政策 2 指出,系统涉及的模型、算法、数据和决策都应被精准记录,以便在发生未知风险时对系统进行审核。同时,人工智能开发者也应使用严谨的科学方法对其算法模型进行反复、定期检测,记录所采用的验证方法和结果并加以公开,用以评估和监测算法模型是否存在潜在威胁。政策 4 认为,应当将安全性放在首要和突出地位。具体来说,在前期部署阶段,应当对人工智能系统进行反复测试,确保人工智能系统与软硬件环境安全交互,依照预期程序稳定运行。同时在系统运行过程中也应当设置必要的监控,根据实际需要随时进行算法的更新、优化和修正等工作。此外,还应确保所有与互联网连接的人工智能系统安全稳定,以免造成网络故障或受到恶意软件感染。

政策 8 从避免人工智能对人类造成伤害的角度阐述了安全原则,认为在人工智能开发与设计过程中,首先要充分考虑其服务或产品可能给人类带来的潜在影响,包括对人工智能使用者的身体或精神伤害、个人信息泄露、个人财产或声誉受损等,确保将所有的伤害降到最低。同时,人工智能开发者可随时进行系统评估,并向公众提供风险评估报告,从而减少潜在的危害行为。政策 7 还进一步指出,开发者需要充分认识人工智能所带来的潜在影响,避免该技术的误用、滥用,将其对人类造成的不利后果的可能性降至最低;此外,提高人工智能系统安全性的关键方法是提高技术的健壮性、可靠性、可控性和成熟度,应当采取必要的安全防范措施,且提高系统运行的可重现性,即验证人工智能系统在相同条件下重复运行时能否表现出相同的行为,从而实现人工智能的系统安全、数据安全和外部环境安全。

5.尊重人的自主性原则

尊重人的自主性(respect for autonomy)原则明确了人工智能对个人自由和自主的尊重,指的是在承认个体所拥有的价值和持有的信念的基础上,赋予具有自主能力的个人以自主选择和自由行动的权力,其内涵包括知情同意、尊重人权与尊严等内容[28]。本研究所选择的 10 个分析样本中,有 4 款人工智能伦理指南规范提到了尊重人的自主性伦理原则。其中,政策 1 和政策 3 提出,人工智能的自主性应当有所限制,要尽可能尊重人的自主性,确保个人可以始终控制人工智能。此外,一是从知情同意的方面来看,政策 4 指出,在开发、设计或使用人工智能系统时,个人必须掌握控制权,即能够随时终止系统算法运行。同时,人工智能系统应具备一定的可扩展性,允许人类随时更改其决策策略,并纳入系统决策支持。二是从尊重人权与尊严的方面来看,政策 7 指出,人工智能系统的设计应当尊重、维系和保护人权,并建议建立统一标准和监督机构,确保人工智能的使用不侵犯人权、尊严、隐私和自由。

6.隐私与个人信息保护原则

隐私与个人信息保护(protection of privacy and personal information)原则规定,人工智能的设计、研发和应用的全流程生命周期都应当保护个人隐私和数据安全。《世界人权宣言》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。”[29] 当下,如何保障隐私安全成为人工智能时代亟待解决的问题。本研究所选择的 10 个分析样本中,有 5 款人工智能伦理规范提到了隐私与保护伦理原则。一方面,人工智能系统可对个人数据进行获取、分析和处理;另一方面,个人也有权保障个人隐私,限制人工智能系统对其个人数据的过度访问、挖掘和利用。具体来说,政策 1、政策 4 和政策 6 指出,人工智能开发者必须安全、合理地使用数据,定期删除无用数据。同时,要设置加密传输和静态数据,通过设置“访问权限”限制特定人员进行数据访问。在收集、使用、共享和存储数据时,人工智能开发者必须遵循必要的隐私政策和个人数据保护条款。政策 8 认为,人工智能开发者应充分保护隐私安全,明确用户数据收集的范围、形式、权利和责任。在使用个人数据时,必须以合法目的为前提,不得侵犯个人或团体的权利。因此,人工智能开发者做好数据隐私安全保护措施,防止未经授权的访问和泄露,建立开放透明的政策和程序,让用户了解其个人数据使用的目的和途径,并为用户提供知情同意许可,及时审查、纠正和删除错误数据。此外,政策 8 指出,当前人工智能开发者经常需要面对各类机密信息,如商业秘密、客户数据、金融信息、研究数据、预发表论文及专利申

请等。这就要求人工智能开发者提高保密意识，除政府等权威机构的特殊要求外，不得随意披露此类信息。

7.透明性原则

透明性（transparency）原则指的是，将人工智能的运作规则、算法参数、决策过程和处理结果以清晰、明确、精准的方式告知公众，主要体现在可追溯性和可解释性两个方面。其中，可追溯性指的是与决策过程相关的数据集的产生过程、处理方法和运行结果能够被精准地跟踪、查询或预测的能力，以帮助人工智能设计者分析、审查其决策的正确性，对系统性能和工作程序做出更好的解释和更优的判断；可解释性指的是对涉及人工智能系统技术研发和应用场景的解释能力，尤其是当人工智能系统对人们的生活产生重大影响时，需要对人工智能系统的决策过程做出适当、及时的解释，从而帮助利益相关者更好地理解系统设计的基本原则、运作原理和适用性。本研究所选择的10款人工智能伦理原则的指南规范中，有7款提到了透明性原则。对于可解释性，政策4认为，人工智能所做出的决策应该具有可解释性，特别是当决策引发公共安全或导致歧视行为时。首先，要确保算法决策的可解释性，即人工智能开发者能够为人工智能系统设计提供充分解释，一些具有潜在公共安全隐患的人工智能系统在发生事故时提供信息支持。其次，人工智能的服务提供商需要为用户提供权限，支持其获取人工智能决策支持的充分解释。对于透明性，政策5指出，首先，人工智能开发者应公开收集得到的数据、数据使用目的，以及在数据处理过程中应用的技术、模型和处理结果；其次，应公开其数据和模型的控制措施，检测是否存在算法歧视，是否需要进行算法修正，实现算法公平。另外，考虑到隐私安全问题，人工智能开发者不可能全面公开其算法源代码，这就需要提高人工智能系统的透明性，明晰责任归属。人工智能开发者还需及时与用户沟通，定期为公众提供系统有效性和可靠运作的监测结果，以帮助公众更好地理解和预测人工智能系统的决策及影响。

8.可靠性原则

可靠性（reliability）原则指的是，开发、部署和设计人工智能系统或服务时，所涉及的算法、模型或数据真实可信，可确保人工智能系统能够在一定时间、一定条件下稳定、高效、无故障地执行指定功能的能力。本研究所选样本中有5款人工智能伦理原则的指南规范提到了可靠性原则。如政策3指出，开发人工智能系统要以大型数据集为基础，应正确理解算法程序涉及的参数和数据，保障数据的完整性，同时要确保开发和改进人工智能系统所需数据的鲁棒性、代表性和可用性。政策6认为，人工智能开发者有义务以科学、严谨、可靠的方式开展人工智能系统的开发和设计，并向用户全面、如实地解释人工智能系统可能存在的技术限制和潜在风险。类似地，政策8指出，人工智能开发者应确保其向使用者提供的系统性能、服务质量、潜在风险等信息都是公开透明的，不得提供虚假或误导性的声明，不得伪造数据资料和实施行贿受贿及其他恶意行为等。

9.其他

除上述伦理原则外，本研究所选择的部分分析样本还提及了其他人工智能伦理原则，用于规范和优化人工智能政策法规、伦理准则的修订与完善。如政策4提出了共同治理原则，认为应当建立更大范围、更加多元的人工智能伦理全球治理机制，即在全球和各区域建立多方利益相关者的人工智能伦理治理机构，邀请民间组织、政府、私营部门、学术界和技术社区及公众共同参与，并依托人工智能伦理问题的举报机制和监督程序，构建科学合理的人工智能伦理安全保护框架。政策6提出了沟通原则，认为人工智能开发者应当与社会公众建立长效沟通机制，尽可能提高用户对人工智能的理解和认识，促进人工智能产品及技术的推广应用。政策7提及了效率伦理原则，并建议开发人员制定相应标准作为有效性的衡量指标，给出各衡量指标的获取方法、解释维度、误差范围、置信区间和结果阐述等，并倡导行业协会和组织为人工智能系统应用的有效性制定测量标准，确保人工智能系统的发展目标、风险评估均符合科学有效的原则；此外，政策7提出了胜任原则，即要求人工智能的开发者和使用者具备人工智能技术有效操作、安全运行所需的知识和技能，并向有关专家和用户提供系统操作方法、前提条

件、操作者技能、系统性能衡量方法以及误操作后果的相关资料。同时,人工智能使用者应当及时接受教育和培训,使其在心理、情感和技能等方面适应人工智能带来的影响,更好地享有人工智能的发展成果。

五、结论与思考

本研究选取了近年来各国学(协)会发布的较有影响力的10款人工智能伦理原则政策文件,试图剖析人工智能伦理原则的核心框架和主要内容。分析认为,现有人工智能伦理原则具有一定的合理性,但仍存在一些不足之处有待改进和完善。特别是,近年来随着人工智能技术日新月异的发展,所开发设计的各类人工智能产品和服务越来越趋向于具有类人的思维和行为,这就要求其应当更加严格地遵守人类基本伦理规范的约束,如家庭伦理原则(包括将夫妻关系与家庭利益相结合,禁止乱伦,对人类自身生产进行规范;要求家庭成员之间相互尊重、相亲相爱、敬老爱幼、无私奉献)[30]、交往伦理原则(即"真"与"诚"原则、仁爱与宽容原则、尊重平等原则)[31]以及职业伦理原则(即自由、责任、诚实、平等、公正和守法)[32]等。但根据笔者研究发现,当前由各国学(协)会发布的人工智能伦理规范并没有涉及有关家庭伦理、交往伦理及职业伦理等层面的基本原则,致使人工智能产品或服务在参与社会实践和人机交往过程中缺乏必要的伦理遵循,进而引发各类社会伦理问题。因此,在制定人工智能伦理原则时,既要考虑为政策执行者提供依据和指导,又要综合分析人工智能产品或服务的现实应用[33],如将真诚原则、仁爱宽容原则、自由守法原则等纳入其中,使其更具普适性和实用性。

参考文献:

[1] KRITTANAWONG C, ZHANG H, WANG Z, et al. Artificial Intelligence in Precision Cardiovascular Medicine[J]. Journal of the American College of Cardiology, 2017(21):2657-2664.

[2] 孙伟平.人工智能是伦理道德建设的"助推器"[J].特区实践与理论,2019(4):98-101.

[3] SWARTE T, BOUFOUS O, ESCALLE P. Artificial Intelligence, Ethics and Human Values: The Cases of Military Drones and Companion Robots[J]. Artificial Life and Robotics,2019(3):291-296.

[4] JOBIN A, IENCA M, VAYENA E. The Global Landscape of AI Ethics Guidelines[J]. Nature Machine Intelligence,2019(9):389-399.

[5] 邓国民,李梅.教育人工智能伦理问题与伦理原则探讨[J].电化教育研究,2020(6):39-45.

[6] 张浩,黄克同.迈向制度化的人工智能伦理[J].人工智能,2019(4):32-38.

[7] 蒋洁.风险视域下人工智能伦理的主旨和原则[J].信息通信技术与政策,2019(6):13-16.

[8] 杜静,黄荣怀,李政璇,等.智能教育时代下人工智能伦理的内涵与建构原则[J].电化教育研究,2019(7):21-29.

[9] 郑志峰.人工智能时代的隐私保护[J].法律科学(西北政法大学学报),2019(2):51-60.

[10] 李伦,孙保学.给人工智能一颗"良芯(良心)"——人工智能伦理研究的四个维度[J].教学与研究,2018(8):72-79.

[11] 杨保军,杜辉.智能新闻:伦理风险·伦理主体·伦理原则[J].西北师大学报(社会科学版),2019(1):27-36.

[12] 李娜,陈君.负责任创新框架下的人工智能伦理问题研究[J].科技管理研究,2020(6):258-264.

[13] HO C W L, ALI J, CAALS K. Ensuring Trustworthy Use of Artificial Intelligence and Big Data Analytics in Health Insurance[J]. Bulletin of the World Health Organization,2020(4):263-269.

[14] 曹建峰,方龄曼.欧盟人工智能伦理与治理的路径及启示[J].人工智能,2019(4):39-47.

[15] 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(13):14-19.

[16] KESKINBORA K H. Medical Ethics Considerations on Artificial Intelligence[J]. Journal of Clinical Neuroscience,2019,64:277-282.

[17] ZENG Y, LU E, HUANGFU C. Linking Artificial Intelligence Principles[C/OL]. Proceedings of the AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019),2018[2021-09-21]. https://arxiv.org/abs/1812.04814.

[18] JESSICA F, ACHTEN N, HILLIGOSS H, et al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI [EB/OL]. [2021-08-23]. https://dash.harvard.edu/handle/1/42160420.

[19] THILO H. The Ethics of AI Ethics: An Evaluation of Guidelines[J]. Minds and Machines,2020(1):99-120.

[20] BURUK B, EKMEKCI P E, ARDA B. A Critical Perspective on Guidelines for Responsible and Trustworthy Artificial Intelligence[J]. Medicine Health Care and Philosophy,2020(3):387-399.
[21] RYAN M, STAHI B C. Artificial Intelligence Ethics Guidelines for Developers and Users: Clarifying Their Content and Normative Implications[J/OL]. Journal of Information, Communication and Ethics in Society, 2021(1):61-68[2021-09-21]. https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html.
[22] Ethical AI Resources. Institutes and Fellowships[EB/OL]. [2021-09-21]. https://ethicalairesources.com/other-resources/.
[23] Future of Life Institute. Summaries of AI Policy Resources[EB/OL]. [2021-09-21]. https://futureoflife.org/ai-policy-resources/.
[24] 准则列表[EB/OL]. [2021-09-21]. http://www.linking-ai-principles.org/cnprinciples.
[25] 斯皮内洛.世纪道德:信息技术的伦理方面[M].北京:中央编译出版社,1998:12-13.
[26] WMA. Declaration of Helsinki[EB/OL].[2021-06-25]. https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/.
[27] 比彻姆.哲学的伦理学[M].雷克勤,郭夏娟,译.北京:中国社会科学出版社,1990:26-27.
[28] KANT I. Foundations of the Metpyysics of Morals[M]. Philadelphia: University of Pennsylvanis Press,1977.
[29] 格德门德尔·阿尔弗雷德松,阿斯布佐恩·艾德.《世界人权宣言》——努力实现的共同标准[M].中国人权研究会,译.成都:四川人民出版社,1999.
[30] 薛宁兰.婚姻家庭法定位及其伦理内涵[J].江淮论坛,2015(6):133-141.
[31] 王倩蒿.论现代社会的交往伦理[D].西安:陕西师范大学,2013:43.
[32] 王海明.制度伦理的重要原则[J].江海学刊,2000(3):89-93.
[33] 刘继虎,王琛扬.智能机器人法律主体地位之探讨[J].重庆邮电大学学报(社会科学版),2021(2):74-80.

The Core Framework and Content Analysis of Ethical Principles of Artificial Intelligence

ZHOU Qingshan[1], ZHENG Xia[2], HUANG Guobin[3]

(1. *Department of Information Management*, *Peking University*, *Beijing* 100871, *China*;
2. *Office of Laboratory Management and Teaching Facilities Development*, *Renmin University of China*, *Beijing* 100086, *China*;
3. *School of Government*, *Beijing Normal University*, *Beijing* 100875, *China*)

Abstract: [Purpose/significance] Artificial intelligence technology quietly opened the curtain of the fourth industrial revolution, bringing revolutionary changes to the development of all walks of life. However, the legal regulations and standard construction of this technology are not yet perfect, which has triggered the global exploration and thinking of artificial intelligence ethics principles and standard systems. National academic/associations have successively issued artificial intelligence ethics codes or policy documents. This research comprehensively analyzes and sorts out the core framework of the current highly recognized artificial intelligence ethical principles, and aims to provide references for countries to formulate and improve artificial intelligence ethical norms. [Method/process] By using the network research method, content analysis method and comparative research method, on the basis of comprehensive consideration of influence, country, availability and other factors, we select the authoritative policy documents related to the ethical principles of artificial intelligence issued by various national academic associations in recent years, and extract the core framework and main contents of the ethical principles of artificial intelligence involved. [Results/conclusion] The field of artificial intelligence mainly includes 8 artificial intelligence ethical principles, including the benefit principle, the responsibility principle, the fairness and justice principle, the safety principle, the respect for human autonomy, the privacy and protection principle, the transparency principle and the reliability principle.

Keywords: artificial intelligence; ethical principles; standard system

(编辑:李春英)