·数字经济与智慧经济·

以人为本的人工智能工程伦理准则探析

□陈光宇　杨欣昱　梁　娜　税发萍　吴　杰　何　甫
[电子科技大学　成都　611731]

[摘　要]　【目的/意义】在快速发展的人工智能给人们带来便利的同时，“去人类中心化”风险的担忧也随之而来。【设计/方法】首先，从技术、利益、责任、环境四个维度开展研究，分析人工智能工程面临的道德风险，若对其不加预防和控制，将有悖于人工智能发展的初衷，即增加人类安全、健康和福祉；其次，通过分类对比分析阿西洛马人工智能原则与欧盟可信赖人工智能伦理准则，阐释工程伦理的三大基本原则对人工智能技术发展和伦理研究的重要性，揭示伦理准则发展和完善的科学路径。【结论/发现】最后，给出“以人为本”的人工智能可持续发展的建议，这些建议对于人工智能工程生命周期中的利益相关者具有重要的指导价值。

[关键词]　人工智能；工程伦理；道德风险；以人为本
[中图分类号] B82　　[文献标识码] A　　[DOI] 10.14071/j.1008-8105(2020)-4011

Analysis of People-oriented Engineering Ethics Principles for Artificial Intelligence

CHEN Guang-yu　YANG Xin-yu　LIANG Na　SHUI Fa-ping　WU Jie　HE Fu
(University of Electronic Science and Technology of China　Chengdu　611731 China)

Abstract　[Purpose/Significance] Along with the convenience brought by the rapid development of artificial intelligence generates widely concerns on the de-humanization hazard. [Design/Methodology] First of all, moral hazard in AI engineering is analyzed from four perspectives, namely, the technology, interest, responsibility, and environment. If the hazard is not prevented and controlled, it may run contrary to the buildup of human safety, health, and well-being, which are the original purpose of AI development. Secondly, through the classification and comparative analysis of the Asilomar AI Principles and the EU' Ethics Guidelines for Trustworthy AI, this paper elaborates the importance of three basic principles of engineering ethics to the development of AI technology and ethical research, thereby revealing the scientific path for the development and improvement of ethics principles. [Findings/Conclusions] Finally, this paper puts forward some suggestions on the people-oriented sustainable development of artificial intelligence, which are of important referential value for stakeholders in AI engineering lifecycle.

Key words　artificial intelligence; engineering ethics; moral hazard; people-oriented

[收稿日期]　2020－08－11
[基金项目]　国家自然科学基金重点项目（71531003）；全国工程硕士专业学位研究生教育在线课程建设项目；电子科技大学研究生精品课程项目；电子科技大学研究生课程思政项目.
[作者简介]　陈光宇（1969–）男，电子科技大学经济与管理学院教授、博士生导师，中国技术经济学会装备质量分会秘书长；杨欣昱（1996–）女，电子科技大学经济与管理学院硕士研究生；税发萍（1995–）女，电子科技大学经济与管理学院博士研究生.
[通信作者]　梁娜（1996–）女，电子科技大学经济与管理学院硕士研究生.

引言

人工智能（AI）是指在机器上实现类似乃至超越人类的感知、认知、行为等智能的系统。人类社会正在由以计算机、通信、互联网、大数据等技术支撑的信息社会，迈向以人工智能为关键支撑的智能社会，人类生产生活以及世界发展格局将由此发生更加深刻的改变[1]。人工智能技术在抗击新冠肺炎疫情中发挥出积极作用，如百度的智能外呼平台用语音机器人及时帮助政府机构快速完成居民排查和通知，阿里安全推出的“AI防疫师”系统提供实时精准测体温、佩戴口罩识别、预警和追踪高危人群等防控功能[2]。人工智能已被视为第四次工业革命的一个标志，具有广泛性的特点，即人工智能应用的场景丰富且多样，不仅在科技领域发展迅猛，更在日常生活领域应用颇多。打败了人类围棋世界冠军的AlphaGo让人们对人工智能有了新的想法，各大网站的智能推荐系统甚至比用户更了解自己，软件的语音识别更加体现了人性化设计。

但个人和社会在享受人工智能技术带来便利的同时，还需要深入考虑人工智能工程的道德风险，即人工智能技术带来的道德结果的不确定性和危害性[3]。在人工智能的设计初期，若不预设相应的伦理道德机制，将会引发极大的不确定性，如微软在2016年3月发布的聊天机器人Tay未设相应的指导机制，在短短24小时之内被互联网网民教成了一个极端主义分子，使得微软不得不将其下线[4]。同样，人工智能在应用中也给个人和社会带来一定的危害，2018年3月，一辆自动驾驶的Uber汽车在美国亚利桑那州撞死行人，公众对于自动驾驶技术的信任危机由此产生[5]，自动驾驶技术也陷入了道德伦理困境，在无法避免事故发生时，是救行人还是救乘客，是救一人还是救几人？随着人工智能应用过程中社会问题的不断凸显，对人工智能工程伦理的思考显得愈发重要，人工智能的研发该设置怎样的伦理边界、秉承怎样的原则、遵守怎样的规章制度，决定着未来人工智能产业的发展方向[6]。

人工智能伦理研究包括人工智能的道德哲学、道德算法、设计伦理、社会伦理四个维度[7]，而人工智能工程伦理更多地关注人工智能研发和使用阶段的伦理原则和准则等问题。近年来，全世界许多国家和国际性组织都先后发布了关于人工智能发展的规划和伦理的一些准则。因此，本文从人工智能道德风险角度出发，阶段性梳理人工智能工程伦理的处理基本原则、《阿西洛马人工智能原则》和《可信赖人工智能伦理准则》等，并进行了对比性分析，提出了相应的建议，将有助于下一步人工智能工程伦理问题的相关制度和法规的完善。

一、人工智能工程的道德风险

人工智能朝着“去人类中心化”方向发展所带来的道德风险值得警惕。未来，基于深度学习的人工智能技术将表现出更快的发展速度、更强大的信息挖掘能力与人机交互能力[8]。董青岭提到的“去人类中心化”概念[9]，表明未来的人工智能不再是人类生产生活的附属品和效率工具的延伸，而是以一种“类人”姿态参与人类社会的运转，机器人网络未来或许会成为社会决策的中心枢纽，人机关系可能会从现在的人机从属关系转变为未来的人机共生甚至是人机相争关系。那么，在人工智能的设计与应用中，我们是否应该放弃人类控制，还是要在多大程度及范围上嵌入人类社会的道德准则呢？

著名科学家霍金指出，人工智能技术创造出的能够独立思考的机器对人类构成极大威胁[10]，因此，为了设计和研发出安全可靠的人工智能系统，我们不仅需要提高技术层面的要求，也需要道德伦理方面的引导和规制[7]。然而，相较我国人工智能技术和产业的发展，相关的伦理理论研究、治理原则和规制规范等配套性制度还比较落后，对于人工智能发展过程中暴露出来的安全问题、伦理问题难以形成通用性的应对方式和解决措施，从而加剧了社会舆论中对此的担忧情绪[11]。

工程含技术要素、利益要素、责任要素、环境要素以及伦理要素，因此人工智能工程伦理问题主要从技术、利益、责任、环境四个方面展开。技术伦理即工程活动技术方面的伦理问题[12]；利益伦理是为争取利益最大化，平衡各方利益、兼顾效率与公平；责任伦理指工程的事前、事后责任以及决策责任与追究性责任引起的问题；环境伦理问题即是工程对环境的影响问题。以下从人工智能工程伦理问题的四个维度讨论其存在的主要的道德风险。

（一）技术维度的道德风险

基于算法和大数据的人工智能技术背后隐藏着风险。其一，算法体现了工程师和程序设计者的逻辑和思想，其政治立场和社会偏见都会不可避免的嵌入算法系统中。其二，人工智能技术涉及大数据，通常存在数据抽样带来的偏差问题，从而在进一步的机器学习的训练中这种偏差无法被消除甚至

被放大。因此，为消除算法与数据抽样带来的偏见问题，设计者与开发者有必要在设计初期就消除其主观偏见，避免不公正的价值观的嵌入，减少人工智能因技术带来的道德风险问题。

《道德机器：如何让机器人明辨是非》一书提到了关于人工智能的道德算法的设计中最典型的三种方法，即自上而下式进路、自下而上式进路和混合式进路。自上而下式进路即理论/规则进路，其本质是“道德规则转换为算法”，在特定的道德场景下提供指导性的抉择标准；自下而上式进路的核心理念是“道德行为来自学习与进化”，即以数据为驱动，利用学习算法使人工智能机器具有道德偏好[7]。但面对的是最新的道德伦理问题且数据抽样有偏差时，不可避免地将产生错误的行为，从而带来负道德效应[9]。近年来，一种称为贝叶斯程序学习（BPL）的新的方法给道德算法的改进提供了新的思路[13]，它是综合运用贝叶斯推理技术和概率生成模型来进行研究。

（二）利益维度的道德风险

人工智能存在侵犯人类利益的风险。首先是安全问题，人工智能工程的安全问题主要指人工智能对于人类来说是安全的、可靠的且不作恶的。自动驾驶技术为作为人工智能工程的典型代表，以减少交通事故为初衷，但近年来发生了几起自动驾驶汽车伤人的事故，引发了社会对其道德风险的思考。其次是隐私问题，人工智能利用大数据技术“了解”用户，掌握了用户的各种信息，在信息高速发展的时代，用户数据存在流入诈骗分子或恐怖分子手中的风险，由此给用户带来隐私被侵犯的问题，甚至面临蒙受生命和财产受损的危险。人工智能工程的基本责任是为人类的生存和发展创造福祉，因此如何通过人工智能工程争取利益最大化，平衡各方利益、兼顾效率与公平是人工智能工程伦理需要解决的核心问题，同时也是衡量人工智能工程好坏的重要标准。

（三）责任维度的道德风险

人工智能在价值选择困境与责任承担困境中存在风险。当人工智能机器像人一样作为自主决策者在遭遇道德两难时应该如何选择，这是在设计之初不得不考虑的问题。譬如无人驾驶汽车在遭遇危险路况时，是救路边的一个人还是救车内的多个人？2016年6月24日《science》上发表了几项关于自动驾驶的研究结果，其中一项就是假如存在一场不能避免的事故，是救车上的乘客还是救路上的行人？根据其调查显示，当车上只有1名乘客并且只有1名行人时，75%的人选择救乘客；当车上只有1名乘客并且有5名行人时，50%的人选择救乘客；当车上只有1名乘客并且有100名行人时，20%的人选择救乘客[14]。在我们生活的真实社会场景中，驾驶员可根据具体的环境和社会的规范约束做出自己的选择，但是对于机器而言，那些预先设定的选择并不能保证伤害最小化。

更重要的是，人工智能机器作为自主行为主体做出判断之后，如何为自己的自主决策承担法律和道义责任。在传统驾驶方式下，交通事故发生后，有较为清晰的法律流程来确定责任主体，并使之与相应的法律责任相匹配。但目前的人工智能机器尚且没有社会规范意识，只能依据内置的算法和道德场景选择判断，没有能力为自己的行为承担相应责任。因此，自动驾驶发展的瓶颈不只在于技术，更多时候是在于道德伦理问题的限制。

（四）环境维度的道德风险

人工智能的发展在一定程度上损害了外部生态环境的利益。因为人工智能工程无法摆脱技术的自然属性，人工智能机器的研制需要从自然界索取不同的物质，从而加速了人类对自然资源的利用和消耗，并且因更新被淘汰的产品其成分在自然界中很难消除。如此一来，人工智能的发展对生态环境会造成很大的影响，可能会使原本严重的生态问题更加恶化，如不加以约束和规范，将对人类追求的环境伦理的可持续性、人与自然和谐相处等基本原则形成巨大的挑战[15]。

因此，若对上述四个维度的道德风险不加预防和控制，人类将面临“去人类中心化”带来的巨大的不确定性，有悖于为了增加人类福祉而发展人工智能的初衷。因此，应坚持“以人为本”制定相应的原则和用于人工智能研发的准则，以支持人工智能稳健、持续地发展。

二、应对人工智能道德风险的基本原则

人工智能工程中面临各种伦理问题，为保证工程的有效开展，保护各方利益，处理人工智能工程伦理问题的三大基本原则显得尤为重要，即“人道主义”“社会公正”以及“人与自然和谐发展”。“人道主义”原则，主要处理工程与人的关系，强调人的重要性；“社会公正”原则是处理工程与社会关系的基本原则，侧重维护社会的公平公正；“人与自然和谐发展”原则，从工程与自然关系的层面，表明工程不能以牺牲自然为代价，做到人与自然和谐发展才是工程的最终目标。这三大原则均

从“以人为本”的观点延伸开来，有助于人们处理人工智能工程中所面临的伦理问题，避免极端主义等有害形式的出现，为世界各国制定相应的人工智能准则提供了理论依据。

（一）“人道主义”原则

人工智能越来越聪明，使得人们不得不提高警惕。如果人们希望人工智能在未来的世界中发挥积极的作用，就必须秉承“人道主义”的原则，以“关怀人类”为宗旨，创造“以人为本”的人工智能社会[16]。首先考虑安全问题。人工智能工程的安全问题主要指人工智能对于人类来说是安全的、可靠的且不作恶的。以自动驾驶技术为例，作为人工智能工程的典型代表，必须保证不伤害人类，时刻将安全问题放在首位。自动驾驶的推广将提高出行效率，解放人类，但公路上的情况错综复杂，一旦有疏忽必将造成重大事故，危及人们的生命安全。其次，还应考虑隐私问题，人工智能工程在应用中势必会涉及个人海量数据，而这就带来了个人隐私和数据安全方面的问题。如手机软件的流行极大地方便了人们的生活，但也为个人数据的泄露埋藏了隐患。因此，人工智能工程必须始终秉承“人道主义”的原则，将人类的人身安全与隐私安全放在首位，确保人们的人身安全及自身利益不受损害。

（二）“社会公正”原则

人工智能研发人员在设计初期应秉承“社会公正”原则，要始终牢记不同区域、不同领域的所有人在人工智能面前是平等的，不应该有人被歧视。首先，在人工智能技术的应用中，为防止种族歧视、资源分配不公等问题的出现，应尽可能消除偏见，其次，还要保护弱势群体。2019年3月，人大代表、科大讯飞董事长刘庆峰表示无论是过去还是未来，作为国内人工智能领域的代表企业，科大讯飞始终着眼于利用人工智能为社会发展赋能，其中最值得关注的便是为弱势群体赋能。人工智能工程在教育上的应用对于缩小教育资源不平衡所带来的差距有着重要的作用，通过“人机共教”“虚拟现实”等技术解决师资水平不平衡和教学设备落后的问题，将有效地保护了弱势群体的利益。人工智能工程目前已从典型应用试点发展到大规模推广阶段，其发展必须时刻秉承“社会公正”的原则，促进社会公平正义，让弱势群体能够享受到更多福利[17]。

（三）“人与自然和谐发展”原则

人与自然和谐发展不仅意味着在工程实践中注重环保、尽量减少对环境的破坏。与其他新兴技术一样，人工智能技术发展的初衷是为了促进人类进步，但是科技的进步往往与资源过度摄取、生态环境破坏有关。电子垃圾的泛滥和“共享单车坟场”给人类敲响了警钟，科学技术的进步不能抛开环境与自然的因素，应足够重视生态系统的完好与多样。人工智能的发展应始终贯穿正义的摄取、确保可持续发展，最终达到人与自然的和谐共生的境地。因此，秉承“人与自然和谐发展”原则尤为重要，人类历史上许多文明皆因生态系统的崩溃，最终导致衰亡。每一项科学技术的发展都应建立在维护“人与自然和谐发展”的基础上，在当今社会大力发展人工智能技术的背景下，秉承“人与自然和谐发展”原则的思想应根植于每一个人的脑海中。

三、典型人工智能伦理准则的对比分析

当前最受关注并得到广泛认可的人工智能准则有2017年由人工智能领域专家联合发布的《阿西洛马人工智能原则》和2019年欧盟委员会发布的《可信赖人工智能伦理准则》，以下通过对比分析，阐释三大基本原则对人工智能技术发展和伦理研究的重要性，以期望找到伦理准则发展和完善的科学路径。

（一）阿西洛马人工智能原则的分析

2017年1月，在美国加州的阿西洛马市举行的Beneficial AI会议上，近千名人工智能和机器人领域的专家联合签署了23条AI发展的原则——阿西洛马人工智能原则（Asilomar AI Principles），呼吁全世界的人工智能开发者在发展AI的同时严格遵守这些原则，以期保障人类未来的利益和安全[18]。“阿西洛马人工智能原则”是20世纪40年代著名的艾萨克•阿西莫夫（Isaac Asimov）的机器人三大法则的扩展版本，主要从科研问题、伦理和价值以及更长期的问题三个方面进行阐述，旨在确保AI为人类利益服务，是当今获得学术界最广泛共识的一项人工智能领域的原则。截至2018年末，已经有3 814人签署，其中包括1 273位来自人工智能和机器人领域的专家学者。

本文着重梳理了“阿西洛马人工智能原则”的第二部分：伦理和价值，该部分涉及条款的6至18条，列举了如安全性、故障透明等13条条款，涵盖了安全职责、技术透明及价值观等方面的内容。通过分析每项具体条款，发现“阿西洛马人工智能原则”的“伦理和价值”部分主要强调人类的重要性与主导性，围绕“人道主义”与“人类中心”两个宗旨，充分体现了处理工程伦理问题中“以人为本”的原则，如表1所示。

表 1 阿西洛马人工智能原则——伦理和价值部分

原则	具体条款	解释
	（6）安全性	AI系统应该是安全可靠的，并接受相关验证
	（7）故障透明性	能及时确定造成AI系统损害的原因
	（8）司法透明性	任何自主系统参与的司法判决都应被相关领域的专家接受
	（9）责任	AI系统的设计开发者有责任和机会去塑造系统的道德影响
	（10）价值归属	AI系统应确保其目标和行为应与人类价值观一致
	（11）人类价值观	AI系统应与人类尊严、权力、自由和文化多样性的理想一致
以人为本	（12）个人隐私	人们应该拥有权力去访问、管理和控制他们产生的数据
	（13）自由与隐私	AI不能无理由地剥夺人们的自由与隐私
	（14）分享利益	AI应惠及和服务尽可能多的人
	（15）共同繁荣	AI应惠及全人类
	（16）人类控制	AI系统应由人类控制
	（17）非颠覆	高级AI绝不能颠覆人类
	（18）人工智能军备竞赛	应该避免致命的自主武器的军备竞赛

（二）可信赖人工智能伦理准则的分析

2019年4月8日，欧盟委员会发布了由欧盟人工智能高级专家组撰写的《可信赖人工智能伦理准则》（Ethics Guidelines for Trustworthy AI）的正式生效文本，为人工智能系统提供具体实施和操作层面的指导[19]，这为欧盟人工智能技术的研发与应用奠定了总体基调，明确提出欧盟各国应协同一致，把握人工智能的发展机遇、应对相应挑战[20]。该准则由欧盟人工智能高级别专家组起草，列出了“可信赖人工智能”的七个关键要求，如图1所示。

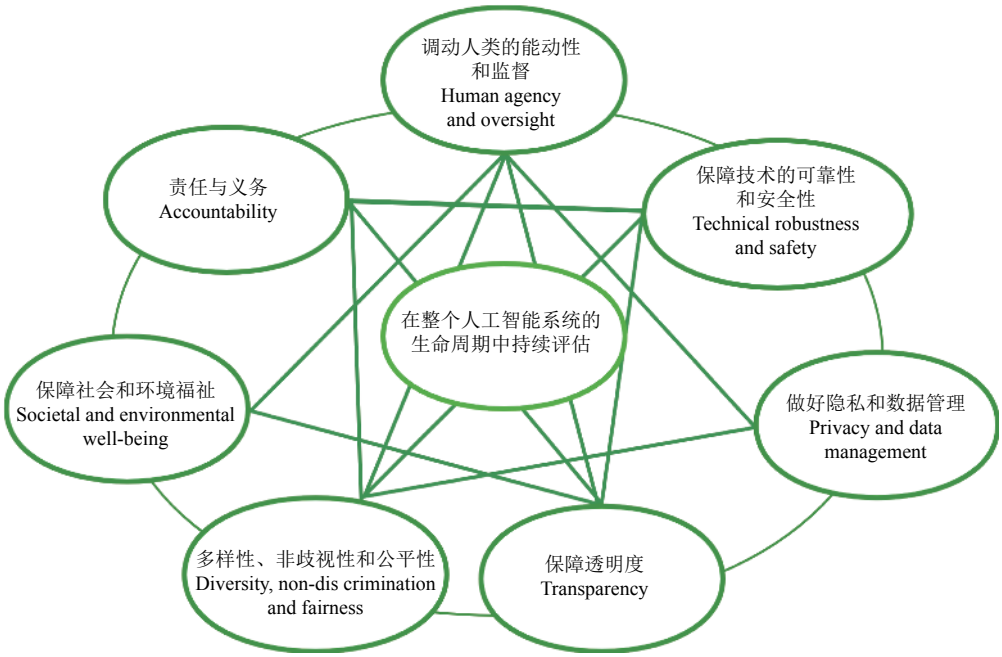

图 1 《可信赖人工智能伦理准则》七个关键要求

这七个措施为世界各国建立安全稳定、可信赖人工智能系统提供了指导方向，规定了人工智能发展道路上必须要遵守的基本要求以及究责机制，实质上，这七个关键要求充分体现了处理工程伦理问题的三大基本原则：“人道主义”“社会公正”“人与自然和谐发展”，如表2所示。

（三）对比分析

1. 共同点

通过对《阿西洛马人工智能原则》与《可信赖人工智能伦理准则》的对比分析，可以发现两项准则都以实现人类利益为出发点，目的都在于维护人类利益不受损害，切实体现了“以人为本”原则。

表 2　《可信赖人工智能伦理准则》——七个关键要求

原则	具体条款	解释
以人为本	（1）调动人类的能动性和监督	依据人类自治原则，AI系统必须支持人类自决，并由人类监督。
	（2）保障技术的可靠性和安全性	首先是保证AI系统可靠，避免黑客利用系统缺陷进行恶意攻击。
	（3）做好隐私和数据管理	隐私权受AI系统影响极大，必须在系统的整个生命周期内确保隐私和数据保护。
	（4）责任与义务	明确AI系统及成果在开发和使用阶段的责任、义务对人类利益至关重要，应遵循公平原则。
社会公正	（5）保障透明度	透明度的要求与可解释性原则密切相关，保证AI系统要素透明度是实现社会公正的基本条件。
	（6）多样性、非歧视性和公平性	社会公正不仅是公平性，还必须在整个AI系统的生命周期中体现包容性、多样性以及非歧视性。
和谐发展人与自然	（7）保障社会和环境福祉	鼓励AI系统的可持续性和生态责任，并将此研究纳入人工智能解决方案以解决全球关注领域。

以人为本，最重要的便是安全问题和隐私问题，在两项准则均得到了充分的体现，如《阿西洛马人工智能原则》中的“安全性”“个人隐私”“自由与隐私”等条款，《可信赖人工智能伦理准则》中的“保障技术的可靠性与安全行”和“做好隐私和数据管理”等。在此基础上，《阿西洛马人工智能原则》进一步围绕“人道主义”与“人类中心”两个宗旨，要求人工智能应惠及和服务尽可能多的人，与人类价值观相一致且完全接受人类控制，绝不颠覆人类。同样，《可信赖人工智能伦理准则》也强调了人类自治的重要性，明确指出人工智能要对人类负责且允许人类监督。由此可以看出，两项准则的共同点都围绕“人类为中心”展开，表明人工智能工程的最基本要求就是站在人类立场上考虑问题，维护人类利益，充分体现了“以人为本”的原则。

2. 各自的侧重点

“阿西洛马人工智能原则”主要围绕“人类中心”进行探讨，更多强调人工智能在维护人类利益上需要遵守的准则，而欧盟制定的《可信赖人工智能伦理准则》不仅强调“以人为本”，又补充说明了公正与生态问题对于人工智能工程的重要性。《可信赖人工智能伦理准则》指出，保障透明度和维护多样性、非歧视性和公平性应贯穿于整个人工智能工程的生命周期。透明度的要求与可解释性原则密切相关，包含与人工智能系统相关的要素透明度，是实现社会公正的基本条件。同时，包容性和多样性以及非歧视性也是实现社会公正的必要条件。另外，准则还指出，在整个人工智能系统的生命周期中，需要保障社会和环境福祉，更广泛的社会、芸芸众生和环境也应被视为利益相关者，人工智能系统需要更注重可持续性并负起生态责任。

与“阿西洛马人工智能原则”不同，《可信赖人工智能伦理准则》的侧重点除了人类，还将公正问题与生态问题纳入考虑范围，不仅秉承了“人道主义”原则，更覆盖了另外两大原则——“社会公正”“人与自然和谐发展”原则。由此可以看出，除了最基本的“人道主义”原则，“社会公正”“人与自然和谐发展”原则也应纳入处理人工智能工程伦理问题的考虑范畴。

通过对上述原则中的13条道德伦理条款和准则中的7个关键要求进行三个基本原则的分类对比分析，得到的共同点和侧重点说明，人工智能准则的发展方向是逐步体现人道主义，到社会和自然公平正义的扩展延伸的发展路径。

四、结论与建议

为避免人工智能“去人类中心化”带来的技术、利益、责任、环境四个方面的道德风险，人工智能的发展应始终围绕“人道主义”“社会公正”以及“人与自然和谐发展”三大基本原则和延伸路径而展开。通过共同点和侧重点分析可以看出，三大基本原则的核心是“以人类为中心”的立场，秉承把人类的安全、健康和福祉放在首位的原则，实现人类社会可持续发展的目标。

上述原则和准则的分析及相关结论对于人工智能工程生命周期中的利益相关者具有重要的指导价值，由此提出进一步的建议：

1. 坚持人类自治原则，强化人类主导和监督。基于人类自治原则，人工智能研发和应用的政策应该将人置于核心，支持人类自治和决策。强化人类主导和监督，人类主导及人为监督有助于确保人工智能系统不会破坏人的自主权或造成其他不利影响，切实保证真正造福于人类、为人类所用。显然，人类应充分发挥自治、主导和监督的作用，对于人工智能工程的成功至关重要。

基于人类自治原则，人工智能研发和应用的政策应该将人置于核心，支持人类自治和决策。强化人类主导和监督，人类主导及人为监督有助于确保人工智能系统不会破坏人的自主权或造成其他不利影响，切实保证真正造福于人类、为人类所用。显然，人类应充分发挥自治、主导和监督的作用，对于人工智能工程的成功至关重要。

2. 促进理论研究，完善制度保障。对于政府，应设立专项资金，支持高校和研究机构开展人工智能等前沿科技的伦理研究，并给予不同人群以学习

了解人工智能的机会，推动全社会对人工智能的知识普及和公共政策讨论。对于社会各界，应展开广泛对话和持续合作，通过一套切实可行的指导原则，鼓励发展“以人为本”的人工智能；并组建由政府部门和行业专家组成的人工智能伦理委员会，对人工智能的开发和应用提供伦理指引，并对具有重大公共影响的人工智能产品进行伦理与合法性评估。

对于政府，应设立专项资金，支持高校和研究机构开展人工智能等前沿科技的伦理研究，并给予不同人群以学习了解人工智能的机会，推动全社会对人工智能的知识普及和公共政策讨论。对于社会各界，应展开广泛对话和持续合作，通过一套切实可行的指导原则，鼓励发展“以人为本”的人工智能；并组建由政府部门和行业专家组成的人工智能伦理委员会，对人工智能的开发和应用提供伦理指引，并对具有重大公共影响的人工智能产品进行伦理与合法性评估。

3. 严格操作规程，落实安全预防措施。从当前技术层面来看，人工智能系统不可信的主要因素之一是发展时间短，技术成熟度不高。因此，为实现可信赖、“以人为本”的人工智能系统，严格操作规程是关键之一。同时要求人工智能工程设计中应考虑可靠的预防性安全措施来防范风险，即尽量减少无意和意外伤害，并防止不可接受的伤害。

从当前技术层面来看，人工智能系统不可信的主要因素之一是发展时间短，技术成熟度不高。因此，为实现可信赖、“以人为本”的人工智能系统，严格操作规程是关键之一。同时要求人工智能工程设计中应考虑可靠的预防性安全措施来防范风险，即尽量减少无意和意外伤害，并防止不可接受的伤害。

4. 加强国际合作，促进共建共享。从人类命运共同体的高度来看，人工智能的发展将在创新治理、可持续发展和全球安全合作三个方面对现行国际秩序产生深刻影响，各国政府与社会各界都应予以关切和回应。各国政府应加强国际合作，建立多层次的国际人工智能治理机制，构建起一套全球性的、共建共享、安全高效、持续发展的人工智能治理新秩序。

从人类命运共同体的高度来看，人工智能的发展将在创新治理、可持续发展和全球安全合作三个方面对现行国际秩序产生深刻影响，各国政府与社会各界都应予以关切和回应。各国政府应加强国际合作，建立多层次的国际人工智能治理机制，构建起一套全球性的、共建共享、安全高效、持续发展的人工智能治理新秩序。

参考文献

[1] 高文, 黄铁军. 从信息社会迈向智能社会[J]. 中国报业, 2020(5): 46-47.

[2] 人工智能在新冠肺炎战“疫”中能帮什么忙? [N]. 人民邮电, 2020-02-25(004).

[3] 常晓亭. 人工智能创新发展中的道德风险研究[J]. 中国医学伦理学, 2020(2): 137-142.

[4] 郭爽. 微软聊天机器人为何会“学坏”[N]. 光明日报, 2016-04-08(010).

[5] 林雨佳. 自动驾驶事故中的过失犯罪分析[J]. 重庆大学学报(社会科学版): 1-10.

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

[7] 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.

[8] 侯佳忆, 吴刚. 人工智能技术发展趋势研究[J]. 科技风, 2020(10): 19.

[9] 董青岭. 人工智能时代的道德风险与机器伦理[J]. 云梦学刊, 2018, 39(5): 39-44.

[10] 霍金. 人工智能可能使人类灭绝[J]. 走向世界, 2015(1): 13.

[11] 王晓丽, 徐鑫钰. 人工智能的道德风险及其治理[J]. 华南理工大学学报(社会科学版), 2020, 22(2): 10-17+2.

[12] 朱海林. 技术伦理、利益伦理与责任伦理——工程伦理的三个基本维度[J]. 科学技术哲学研究, 2010, 27(6): 61-64.

[13] LAKE B M, SALAKHUTDINOV R, TENENBAUM J B . Human-level Concept Learning Through Probabilistic Program Induction[J]. Science, 2015, 350(6266): 1332-1338.

[14] BONNEFON J F, SHARIFF A, RAHWAN I. The Social Dilemma of Autonomous Vehicles[J]. Science, 2016, 352(6293): 1573-1576.

[15] 陈首珠. 人工智能技术的环境伦理问题及其对策[J]. 科技传播, 2019, 11(11): 138-140.

[16] 李真真, 齐昆鹏. 人工智能——“以人为本”的设计和创造[J]. 科技中国, 2018(3): 94-97.

[17] 张德帅. 探析“人工智能+教育”对我国教育资源公平分配的促进作用[J]. 中国新通信, 2018, 20(19): 191.

[18] 阿西洛马人工智能原则——马斯克、戴米斯·哈萨比斯等确认的23个原则,将使AI更安全和道德[J]. 智能机器人, 2017(1):20-21.

[19] Ethics Guidelines for Trustworthy AI[EB/OL]. (2019-06-27). https://apo.org.au/node/244266

[20] 殷佳章, 房乐宪. 欧盟人工智能战略框架下的伦理准则及其国际含义[J]. 国际论坛, 2020(2): 18-30+155-156.

编 辑 邓婧