【伦理与道德】

人工智能的伦理风险治理探析*

张　铓

摘　要:人工智能技术增进社会福祉和进步的同时,伴随着人机边界模糊、责任追究困境、价值鸿沟扩大和技术过度依赖等伦理风险。人工智能伦理风险的生成有其内在逻辑。人工智能伦理风险的深层性、复杂性等特征契合了协同治理的要求。协同治理范式是人工智能伦理风险治理具有可行性的新探索。推进人工智能伦理风险的协同治理,应构建多元协同组织,形成以政府为核心,技术专家、社会组织、研究机构和公众等共同参与的风险治理自组织系统。在此基础上,通过夯实协同治理条件、增强协同治理参量提升人工智能伦理风险治理效能,促进人工智能伦理风险的整体性治理。

关键词:人工智能;伦理风险;协同治理
中图分类号:B82　　**文献标识码**:A　　**文章编号**:1003-0751(2022)01-0114-05

近年,人工智能技术发展迅猛,掀起和引领着第四次工业革命,正深刻改变社会、改变世界,人工智能时代已经提前到来。新冠肺炎疫情期间,从病毒分析、疫苗开发、药物研发到诊断辅助、智能测温、AI消毒,人们已经能感受到人工智能技术给社会生产和生活方式带来的广泛影响。然而,与此同时,人工智能技术的复杂性和不确定性也带来了伦理风险。作为一种具有颠覆性影响的技术,人工智能的伦理风险开始显现。面对人工智能的伦理风险挑战,不同的学者基于不同的视角进行了分析和探讨,推动了人工智能伦理风险治理的研究,但仍有一些值得商榷之处。一是对人工智能伦理风险的界定不够清晰,有的研究将人工智能的"一般风险"视为"伦理风险";二是对人工智能伦理风险形成机理的揭示有待深化;三是对人工智能伦理风险治理对策的系统性把握有待加强。为此,本文聚焦人工智能的伦理风险,拟在描绘人工智能伦理风险镜像的基础上,分析人工智能伦理风险的生成逻辑,探讨人工智能伦理风险的协同治理之道。

一、人工智能伦理风险的镜像素描

人工智能技术的兴起不仅是技术和产业领域的重大革命,更引发了社会治理等众多领域的深刻变革,带来了不可忽视的伦理风险。了解人工智能技术的伦理风险是规范人工智能的研发和应用、有效治理其伦理风险的前提,亦是人工智能技术可持续发展的基础。

1.人机边界模糊:伦理关系失调风险

人工智能技术的快速发展和不断突破,有朝一日,有可能研发出具有自我意识的智能产品,人机边界将变得模糊。那么届时,人与有自我意识的智能产品之间是什么关系?此种智能产品是否应具有道德主体地位?由此,引发了"心智边界之争"。面对该问题,伦理学者的解答不尽相同。人类中心主义者认为,道德主体资格是人类独有的,人的道德地位是基于人类的精神特点存在的,智能机器是人创造

收稿日期:2021-12-03
*基金项目:国家社会科学基金重大专项课题"建立健全我国网络综合治理体系研究"(20ZDA062);浙江省哲学社会科学规划课题"抗疫彰显的中国制度优势研究"(21GXSZ019YB)。
作者简介:张铓,男,浙江工商大学公共管理学院副教授(杭州　310018)。

的高级工具和设备而已，“它们的能力不过是在例示过程中随着计算机的运行而产生下一步的形式体系”①。持此类观点的学者不赞成智能机器具有道德主体资格。也有一些学者要求重新定位人工智能产品的道德地位，认为人类之外的生命体具有道德主体资格的可能性。②如果赋予智能机器人道德主体资格，智能机器人参与到人类的政治、经济和社会活动中，那么，该如何界定人与智能机器人的关系呢？智能机器人具有道德主体资格的同时，能承担相应的道德责任吗？总之，人工智能时代如何界定人和智能机器的关系是一大难题。人工智能的发展将挑战人类文明已经形成的伦理关系，引发新的道德冲突和伦理困境。

2.责任追究困境：伦理规范失效风险

人工智能技术的算法生产和运作过程的“黑箱化”不仅对有效监管造成困难，而且因为其在实际应用中往往居于“幕后”，其运行过程的不透明和治理规则的不易解释，造成责任追究的伦理难题。英国《卫报》专栏作家 Ben Goldacre 将算法比喻为一系列的“黑盒子”。面对这样一个知之甚少的数字世界，“人类并不清楚什么是安全的、什么是有风险的”③。需要指出的是，人工智能技术环境中治理规则的难以解释并非完全因为公司或个人的隐瞒，其深刻原因在于算法和机器学习的复杂性。进入人工智能时代，人们发现传统的技术责任伦理和制度规范已无法有效应对智能技术应用中的风险责任界定和追究。基于传统的技术责任伦理和制度规范，设计者应承担技术风险的主要责任。然而，人工智能时代如果将智能机器导致的侵权行为完全归责于设计者，则并不具有说服力。倘若归责人工智能本身，又该如何问责一个机器呢？由此，长久以来基于行为和后果之因果关系的技术责任伦理和法律制度，遇到严重挑战。简言之，责任追究困境问题是人工智能技术的又一伦理风险。

3.价值鸿沟扩大：伦理价值失衡风险

人工智能时代权力的性质及其实现形式发生了变化。人工智能技术嵌入社会系统中，一方面，人工智能技术的算法催生出隐形权力。算法黑箱为权力的隐性运行提供了条件。算法作为以计算机代码表达的意见，算法的设计、目的、标准必然体现设计者和开发者的主观意志和价值取向。算法的设计者和开发者很可能将其主观偏见带入算法系统，从而形成算法偏见、算法歧视，导致种族、性别、年龄和阶层区域歧视，使得社会运行过程中缺失公共价值，技术的工具理性与价值理性失衡。例如，弗吉尼亚·尤班克斯（Virginia Eubanks）指出，大数据与算法应用于穷人和工人阶层，“强化和延续了近代的济贫措施中的道德主义倾向和惩罚性后果”④。另一方面，人工智能技术通过重构产业结构和社会分工模式，可能引发结构性失业，危及社会公平和正义。智能化、高效化的人工智能机器不仅会替代人做“脏累差”的工作，甚至可能取代一些职业。虽然人工智能的发展会创造新的就业岗位，但并非所有人都能跨越技术鸿沟，适应人工智能时代的工作变革。

4.技术过度依赖：伦理行为异化风险

人工智能技术的应用给人类生活带来便利的同时，也可能造成对智能技术的过度依赖，导致伦理行为异化。一方面，智能机器的精准化、个性化信息推送可能削弱人的甄别能力，强化人的行为偏见，使人盲从某些错误观点。在社交媒体中，应用智能机器可分析人的观点、兴趣爱好等，然后根据这些数据分析实现个性推荐，造成人们的行为偏见和甄别能力削弱。另一方面，人工智能在社会系统中的应用可能造成人对技术的严重依赖，甚至社会治理的角色替代。对技术治理的迷信和过度依赖会造成社会治理创新不足、缺乏实质增长的“内卷化”现象。例如，有关部门对信息技术资源不断投入的同时，条块结构越来越复杂，对技术工具的依赖也越来越大。严重依赖技术工具的社会治理模式势必会逐渐消解治理主体的主动意识和创新精神，造成智能技术的算法“决策”代替治理主体的自主决策。在此境遇下，人不再是治理主体，反而可能成为算法权力中被计算的客体。当前虽然人工智能技术总体上处于“弱人工智能”阶段，智能机器尚不具备社会治理主体资格，但是，随着“强人工智能”和“超人工智能”的开发和应用，社会治理主体有可能被人工智能取代，这并非危言耸听。

二、人工智能伦理风险的生成逻辑

技术是一把“双刃剑”。作为一种新兴技术，人工智能技术在增进社会福祉和社会进步的同时，其潜藏的伦理风险也逐渐显现。人工智能伦理风险的生成遵循技术演变的规律，有其内在逻辑。揭示人工智能伦理风险的生成逻辑，可为有效治理其风险

提供科学依据。

1.内源性逻辑：算法黑箱与智能技术的不确定性

技术本质是一种解蔽方式，然而，解蔽的过程暗藏风险。人工智能技术的算法背后隐藏了“技术黑箱”，人们在应用人工智能技术的过程中有可能被技术“绑架”。算法是人工智能技术的基础，开发者有可能将其自身所持有的价值偏见植入算法中，甚至存在非法篡改算法参数的风险。⑤算法的本质是一系列指令，这些指令很难转化为通俗易懂的语言。一方面，人们很难理解、预测和评估人工智能技术背后的决策逻辑。近些年，人工智能与大数据、物联网、云计算和脑科学等技术深度结合后，其背后的算法逻辑更加复杂，导致技术系统的潜在风险扩大。另一方面，人工智能技术的道德算法并非完全独立运行。人工智能的算法运作具有高度的复杂性，其依存的道德算法需与其他系统结合完成深度学习，才能发挥应有的作用。换言之，人工智能技术的道德算法不是一个独立的运作体系，其依赖于代码和数据样本的机器学习具有较大的不确定性。多次重复使用的样本数据容易误导智能机器，导致智能机器做出错误的道德决策，人工智能的伦理风险由此产生。

2.功能性逻辑：技术赋权与技术约束的双重效应

技术并不是价值无涉的，人工智能技术在赋能社会和公民的同时，也呈现出技术约束性的“另一面”。当人工智能技术与公共权力结合时，如果公共权力使用不合理，人工智能技术的约束效应就会被放大。例如，维护安全稳定是社会治理的重要目标之一，掌握公共权力的政府部门在应用人工智能技术方面具有显著优势，然而，个别政府部门会以公共安全为由过度收集和使用公民信息，不合理使用人工智能技术工具，对公民隐私权利造成侵犯。作为一种治理技术，人们在应用人工智能技术的过程中应重视技术价值理性的发挥。如果过多关注人工智能技术的工具理性而忽视其价值理性和人文关怀，那么，人工智能技术的“异化”和伦理风险的生成就不可避免。进而言之，人工智能的发展不应脱离伦理道德和制度规范的约束。当前，传统的技术伦理道德和制度难以适应新兴技术的发展。面对人工智能技术的迅猛发展态势，必须建立和完善新的技术伦理道德和制度，以规约人工智能的伦理风险。总之，人工智能的发展和应用需坚持以人为本，“重视对人的尊严、自由和社会民主、平等、公正等重要价值的维护”⑥，保持技术工具理性和价值理性的平衡。

3.现实性逻辑：风险研判与防控能力的相对不足

人工智能的伦理风险是主观认知性和客观实体性的结合，前者主要指人类对其伦理风险的认知心理、感知判断等主观元素，后者主要指人工智能技术自身存在的复杂性、深层性和不确定性等客观现实。然而，囿于现实条件，人类对人工智能技术风险的研判和防控能力有一定的局限。一方面，作为复杂的技术系统之一，人工智能技术背后的算法数据极为庞杂，当前人类尚未能有效解决算法逻辑的不确定性，因此无法全面预测人工智能的潜在风险。另一方面，对智能技术认知的有限性影响公众应对人工智能伦理风险的能力。这是人工智能伦理风险产生的现实原因。近些年，人工智能的快速发展和深度应用已带来一系列伦理风险问题，对人类的法律制度、伦理规则和技术规范产生挑战。例如，智能机器是否具有伦理道德主体地位，人工智能技术应用导致的法律责任如何界定，人工智能技术发展带来的价值鸿沟失衡该如何弥合。然而，相较于发展迅速的人工智能技术，人类社会应对其伦理风险的研判和防控能力则稍显薄弱，还亟待提升。

三、人工智能伦理风险的协同治理

面对人工智能技术带来的伦理风险和挑战，众多国家提出了治理措施，但无论是“无须批准式监管”还是“审慎监管”治理路径，仍然停留在原则、准则和战略框架层面。⑦实践证明，人工智能的伦理风险具有复杂性和深层性，仅凭政府一方难以有效治理。再则，人工智能的发展和应用涉及多元利益主体，因此对其伦理风险的治理亦要求多元主体参与共治，正如“技术治理需要利用自由、分权和参与营造的社会空间”⑧，技术风险的治理也应倡导社会合作，而不是由政府单一治理。人工智能伦理风险的治理应构建一种组织网络，通过协商合作重构治理主体间关系，以超越传统的技术风险治理模式。人工智能伦理风险的深层性、复杂性等特征契合了协同治理的内在要求。协同治理范式的理论基础是协

同理论,协同理论强调不同子系统以目标共识为基础,通过互动和合作推进整体性治理。协同治理范式是人工智能伦理风险治理具有可行性的新探索。

1.构建人工智能伦理风险协同治理组织

对人工智能伦理风险的治理是政府的职能之一。人工智能的伦理风险给现代政府管理带来了挑战。政府拥有技术风险治理方面的资源和权力优势,是人工智能伦理风险治理的核心主体。例如,2019 年,美国启动“AI 计划”,成立国家标准技术研究院(NIST)等政府机构规范人工智能产业发展。然而,由于人工智能的伦理风险涉及面广,其治理具有高度的复杂性,政府单一主体难以实现有效治理。人工智能伦理风险的有效治理应发挥多元社会主体作用,搭建合作共治框架。构建多元协同组织是人工智能伦理风险协同治理的基础。为此,要以制度规范明确多元治理主体在人工智能伦理风险治理中的权利和义务,实现政府单一治理向多元共治模式的转变,构建政府、技术专家、社会组织、研究机构和公众等共同参与的技术风险治理组织网络。该组织网络有助于克服单一主体治理人工智能技术风险的局限性,形成合作共治的风险治理格局。在具体实践中,要打破传统“命令—控制”式的技术风险治理范式,倡导政府与技术专家、社会组织、研究机构和公众的互动交流和民主协商,形成风险治理主体之间紧密的合作关系,共同防范和化解人工智能伦理风险。要畅通社会参与机制,积极引导多元社会主体参与人工智能伦理风险治理规则的制定,必要时可引入第三方监管机构。总之,构建多元协同组织,实现人工智能伦理风险的协同治理,应形成以政府为核心,技术专家、社会组织、研究机构和公众等共同参与的风险治理自组织系统。需要指出的是,构建协同组织仅是人工智能伦理风险协同治理的必要条件之一,要实现人工智能伦理风险的协同治理,还需夯实协同治理条件,增强协同治理参量。

2.夯实人工智能伦理风险协同治理条件

夯实协同治理条件有助于提升人工智能伦理风险治理的效能。具体而言,要完善四个层面的协同治理条件。一是提高技术伦理风险认识。面对人工智能技术的飞速发展,一种过于乐观的观点认为,技术发展本身孕育着技术风险的解决方案。因此,目前遇到的人工智能的伦理风险治理难题,在未来自然会解决。核技术发展等事实充分证明,这种对于技术发展的盲目乐观和过度自信并不可取。因此,加强人工智能伦理风险认知教育很有必要。一方面,要面向社会开展人工智能技术科普活动,让公众更好地了解人工智能技术的发展趋势及其潜在风险;另一方面,要加强人工智能研究共同体的责任伦理教育,对人工智能研发人员开展科技法律和科研伦理方面的宣传教育,增强其技术风险责任意识,提升人工智能研究共同体的社会责任感,使其做到“负责任创新”地开展人工智能研究。二是强化技术伦理规范建设。当前,世界上掌握人工智能技术的主要国家相继成立人工智能伦理委员会,制定相应伦理标准,规范人工智能伦理风险。例如,2018 年,欧盟发布了人工智能伦理准则。我国应结合实际,借鉴国外先进经验与做法,成立多方参与的人工智能伦理委员会,进一步完善人工智能伦理准则和道德规范,强化负责任的智能技术创新。三是完善相关法律法规。一方面,要厘清智能机器人的法律主体地位,充分讨论和论证智能机器人法律人格创设的可能性,健全人工智能研发和应用的问责机制;另一方面,要完善人工智能产业发展相关法律法规,细化人工智能研发、市场等准入规范,明确人工智能开发主体的法律责任。在发挥市场竞争和市场机制优胜劣汰作用的同时,加强人工智能产业的制度规范。政府部门要成立相关监管机构加强对算法的治理与监管,彰显算法正义,避免人工智能技术应用被“资本”绑架,促进人工智能产业的健康有序发展。四是加大社会政策支持。人工智能技术的发展将导致就业结构、就业方式的深刻变革,因此,在人工智能发展相关公共政策的考量中,不仅要关注人工智能发展的产业和经济政策,也要关注人工智能发展的社会政策。要坚持以人为本,完善社会保障制度,建立终身学习和就业培训体系,有效应对人工智能发展可能带来的结构性失业等社会问题。

3.增强人工智能伦理风险协同治理参量

增强协同治理参量对促进人工智能伦理风险整体性治理具有积极作用。一是完善技术风险治理协同机制。人工智能伦理风险的治理涉及多方利益。因此,要构建技术风险治理的利益表达、补偿及协调机制,搭建技术风险信息共享平台,促进技术风险信息分享。二是建立技术风险评估机制。人工智能技术发展具有较大的不确定性,因此需要运用多种手段和工具,对其伦理风险进行科学评估,开展人工智

能技术伦理风险预警。一方面,政府可联合高校、智库和企业等对人工智能技术进行安全等级评估,有针对性地采用安全防护技术,规避人工智能技术可能出现的故障、失控和被入侵等风险,确保人工智能技术在安全、可控的范围内发展和运用。另一方面,要积极开展人工智能社会实验研究。积极开展人工智能社会实验研究有助于科学研判人工智能技术的社会风险,“准确识别人工智能对人类社会带来的挑战和冲击,深入理解人工智能的社会影响特征与态势,深刻把握人工智能时代社会演进的规律”⑨。三是构建技术风险治理共同体。作为全球性技术,人工智能的伦理风险治理是跨国家和地区的。由于不同国家和地区之间技术标准、准入制度等不尽相同,人工智能风险治理的体制和机制也存在差异。因此,要在人类命运共同体理念的指引下,建立和完善跨国家和地区的技术风险治理协调和联动机制,形成技术风险治理共同体,促进人工智能风险的协同治理。

综上所述,人工智能技术是一种尚未成熟的革命性技术。人工智能技术在增进社会福祉和社会进步的同时,伴随着人机边界模糊、责任追究困难、价值鸿沟扩大和技术过度依赖等伦理风险。对此,要统筹兼顾,多维并举,通过协同治理有效防范和化解其风险。随着数据驱动算法的升级和优化,人工智能技术的发展和应用具有广阔的前景。我们要因势而为,在人工智能伦理风险的协同治理中构建技术与人、技术与社会的和谐生态,推动社会治理变革和创新,促进国家治理体系与治理能力现代化。

注释

①Searle JR. Minds, Brains and Programs. *Behaviroral & Brain Sciences*, 1980, Vol.3, No.3, p.417.②闫坤如、马少卿:《人工智能伦理问题及其规约之径》,《东北大学学报》(社会科学版)2018年第7期。③Goldacre, B. When Data Gets Creepy: The Secrets We don't Realise We're Giving Away, *The Guardian*, 2014-12-05.④Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. St. Martin's Press, 2018, p.37.⑤谭九生、杨建武:《智能时代技术治理的价值悖论及其消解》,《电子政务》2020年第9期。⑥张铤、程乐:《技术治理的风险及其化解》,《自然辩证法研究》2020年第10期。⑦贾开、蒋余浩:《人工智能治理的三个基本问题:技术逻辑、风险挑战与公共政策选择》,《中国行政管理》2017年第10期。⑧彭亚平:《照看社会:技术治理的思想素描》,《社会学研究》2020年第6期。⑨苏竣、黄萃:《探索人工智能社会治理的中国方案》,《光明日报》2019年12月26日。

责任编辑:思　齐

An Analysis on Ethics Risk Governance of Artificial Intelligence

Zhang Ting

Abstract: While artificial intelligence promotes social welfare and progress, it also brings about ethics risks such as the blurred boundary between man and machine, the dilemma of blaming responsibility, the larger gap in moral values and over-dependence on technology.The generation of ethics risk of artificial intelligence has its internal logic. The deep and complex characteristics of ethics risk of artificial intelligence meet the requirements of collaborative governance. The collaborative governance model is the new plausible exploration of artificial intelligence ethics risk governance. To promote this, multiple collaborative organizations should be established to form a self-organizing system of artificial intelligence risk governance with the government as the core and the participation of technical experts, social organizations, research institutions and the public. On this basis, by optimizing the collaborative governance conditions and enhancing the collaborative governance parameters, we can improve the governance efficiency of the ethics risk of artificial intelligence and promote the collaborative governance of the ethics risk of artificial intelligence.

Key words: artificial intelligence; ethics risk; collaborative governance