

人工智能技术应用风险与规避思路

杨莉, 刘文文*

(兰州理工大学马克思主义学院, 甘肃兰州 730050)

摘要: 随着人工智能技术的快速研发和广泛应用, 加强对人工智能技术的伦理风险预估和政策规避路径研究已然成为全球治理共识。为此, 应当重新审视人工智能技术本质, 正视人工智能技术应用带来的诸多技术风险、伦理风险、社会风险和法律风险挑战。实际上, 人工智能技术嵌入实际应用风险的规避是一项复杂的系统工程, 应该致力于构建一个综合性的治理框架。简而言之, 应强化风险意识, 制定技术伦理规范和培养科学家的人文精神, 提高人工智能技术模型的可解释性和可追溯性; 以社会合作共治为框架, 推进人工智能技术风险的协同治理; 推动人工智能技术运用的国际合作, 以及人工智能技术伦理与法律标准化治理体系建设。

关键词: 人工智能技术, 风险, 规避路径

进入21世纪以来, 以人工智能 (Artificial Intelligence, 简称AI) 技术为基础的新一代信息技术不断向认知智能阶段迈进, 特别是2022年11月美国人工智能Open AI公司研究发布的聊天机器人程序ChatGPT (Chat Generative Pre-trained Transformer), 实现了人工智能从感知理解世界到生成创造世界的跃迁。ChatGPT凭借强大的扩展迁移能力不断渗透到人类社会和人本身的各个领域, “隐而不彰地重塑社会结构和治理形态”^[1]。然而, 人工智能技术既是社会和人本身发展的强大推动力, 创造诸多新型领域的成就和便利, 也暴露出内在固有的不确定性风险和挑战, 如技术工具性风险、劳动风险、隐私泄露风险、算法歧视风险等, 有关人工智能威胁论的话语不断出现在社会生活中。当前, 人工智能技术应用风险的研究已然成为全球治理共识, 国内外学者分别从不同角度对人工智能技术展开了研究。国外学者Cath Corinne^[2]对人工智能技术的

研究与应用聚焦于技术自身所引发的治理问题, 如算法治理、技术伦理、技术监管等; Justin B Bullock^[3]学者指出, 伴随着人工智能在治理和社会中的日益普及, 可以将人工智能作为社会治理工具。国内学者张欣^[4]认为, 未来还需从监管基础设施、监管工具体系、监管架构变革以及监管制度环境等核心要素出发, 加以持续不断的探索与优化人工智能技术治理方式。张辉^[5]等学者认为, 人工智能治理需要维护公民精神上的尊严与平等, 防止人工智能技术带来性别、种族、职业、年龄、地域、学历和兴趣爱好等方面的歧视, 带来精神伤害和权利损失, 促进社会公平。

综上所述, 目前国内外对人工智能技术风险和治理研究较为丰富, 但国内外学者研究深入挖掘这些问题产生的深层次原因尚不多见。同时, 对于人工智能技术本质、人工智能技术风险类型、为什么进行治理以及如何治理等还没有形成深刻的成果与共识, 更没有整体性的认知和分析

作者简介: 杨莉, 兰州理工大学马克思主义学院, 教授, 硕士生导师, 研究方向为马克思主义理论, 科学与社会。

刘文文, 兰州理工大学马克思主义学院, 研究方向为马克思主义基本原理。* 通讯作者。

框架，制约了人工智能技术治理的发展。因此，在我国大力建设科技强国的现实需要，深入分析人工智能技术本质和技术应用的风险类型，对于建构人工智能技术应用风险治理综合性分析框架，指导其治理体系和治理能力的现代化实践具有深远的现实意义。

1. 人工智能技术的本质

进入智能化、信息化时代，人工智能技术具有不同于以往任何科技的特殊性，试图混淆人的本质和主体性，成为束缚人类实现自由而全面发展的重大阻碍。人工智能技术本应成为促进社会进步和人类实现“类本质”的强有力技术手段，似乎凭其一己之力使人类面临一系列严峻挑战。以往与人工智能技术相关的“技术全能论”“技术天然合理论”“技术完全自主论”“技术至上论”等代表性观点，主要聚焦于形而上学的唯技术论，未必能够正确反映人工智能技术的本质。在马克思看来，即便人类的存在形态与社会生产方式密切联系，并且生产工具从手工形态更替到人工智能逐渐具备代替人脑的趋势，但它实际上难以获得人类主体性意识，依旧是人类智能发展的产物。

首先，人工智能技术本质是人类社会实践活动的阶段产物，人类智能是人工智能存在的先决条件。人工智能是人类智能创造的、可控的机器或技术，它仅仅发挥了人类本性中学习、认知、模仿和复制的一个方面。从根本上区别人工智能首先需要充分理解什么是人类智能，这是人工智能不同于人类智能的特殊性所在。人类智能以大脑为核心依赖于复杂的生命系统，结构视觉、数字记忆、分析思考、创意思维和语言交流等人类特有的部分能力。马克思从实践哲学视角探寻人类的逻辑起点在于“现实的人”的本质及其存在状态，人的本质属性就是人类在现实物质生产活动中不断生成的错综复杂的社会关系的

总和，指向人类自由解放的宏伟目标。马克思认为：“人的本质不是单个人所固有的抽象物，在其现实性上，它是一切社会关系的总和”^[6]，人类实现“类本质”是在不断地进行社会批判过程中实现的，也就是在“改变世界”的否定性中自我生成的。人的本质存在方式就是自由自觉的社会实践活动。而人工智能技术不具备人类的本质属性，因而无法真正地同人类进行自由自觉地社会实践和交往活动。就人类整体而言，人工智能无法脱离人类智能存在。人工智能基于语言符号系统，从根本上依附于人类自身，它的诞生是人类为了突破个体的局限性和脆弱性而研发和运用的技术。虽然人工智能技术同人类智能一样都有从事社会物质生产实践的能力，仿佛可以脱离人类智能具备了自己的独立智能，但人工智能实际上没有完整意义上的自我意识，无法拥有人类特有的理想信念、情感意志等意识形态内容，只能通过人类事先预设好的程序进行机械的逻辑运算，是一种横向信息传递的反应特性。也就是说，人工智能“虽然有智能却没有智慧，虽有智商却没有情商，虽会计算但不会算计，虽是专才却并非通才”^[7]，并没有脱离人类智能，获得自主能力和独立意识。所以，人工智能只是人类漫长社会发展阶段的创造性产物，是人类智能的一种表现形式，从根本上无法超越人类智能。

其次，人工智能是人类逻辑思维实体化的主要表现方式。人工智能技术的发展经历了机械化向智能化的转变过程。在帕斯卡的基础设计上，莱布尼茨为了实现逻辑运算发明了二进制计算法则，为现代计算机奠定了科技基础。人工智能借助二进制法则和编译建构模型实现了对人类逻辑思维的模拟，具备了单一的逻辑思维能力。但人工智能实际是依靠逻辑运算对人类思维进行模拟的计算机程序，是人的本质力量的对象化，不存在与人类一致的实践关系和交往关系，人工智能本质是对部分人类智能的模拟、替代或者超越。

通常对人工智能研究领域的定义主要包含五层框架, 从底层到顶端分别是: 数据和计算能力的基础设施建设、以LSTM序列学习和深度学习为主的机器学习算法、计算机视觉和语言处理能力为主的技术方向和问题、图像识别和机器翻译为主的运用技术以及金融和医疗为主的行业解决方案。可见, 人工智能的核心依旧是数据、算法和计算能力, 不论人工智能以何种范式出现, 它都只是一种逻辑运算程序。人工智能发展经历了符号主义、联结主义和行为主义的三大智能路线: 符号主义探索人工智能的方式是符号的转换和操作, 主张用公理和逻辑体系搭建一套人工智能系统; 联结主义探索人工智能的方式是对大脑的仿真模拟技术, 它的研究者们建议从观察人类大脑运行的生理学机制出发, 模拟人类神经网络建立人工神经网络以达到模拟人类智能的目的; 行为主义通过构建适应于环境的行动模块来提升机器人的智能, 即人工智能在复杂环境下的行为控制问题。虽然三大研究范式可以相互联系和补充, 但它们实际上都是对人类智能的侧面模拟。

最后, 人工智能技术是人类劳动力的延伸, 仍然是人类社会的生产工具。马克思指出劳动工具“是人的手创造出来的人脑的器官; 是对象化的知识力量”^[8], 人类使用劳动工具的目的就是更加积极主动地认识世界和改造世界。人类作为历史活动的主体和最终目的, 探索人工智能技术的目的就是不断满足认识世界和改造世界的要求, 它是人类赋予计算机某种智能帮助人类从繁重的劳动中解放出来的新型劳动生产工具, 服务于社会生产力和社会关系解放目的。回顾人类社会历史和劳动工具变革史, 人类最初发明石器、铁器工具仅仅是为了满足在原始社会恶劣环境生存的需要。人类在对社会生产实践不断进行自我否定、超越、创造中超越落后生产力的局限。在工业资本主义时期, 由锭子、箭头、针等工具构成的纺纱机器生产取代了工场手工业, 代替了工人

使用工具的技能, 促进了生产力的大幅发展和资本主义文明的出现。人类逐渐意识到科学技术的重要性, 不断地发明新的生产工具以达到促进社会生产力进一步解放的目的。计算机网络技术是人类技术发明的最新形态, 科技以一种更高级的样态呈现在人类社会, 劳动工具的功能不断超越人体器官的劳动能力, 呈现出不断替代人类身体器官的趋势。从这种维度出发人工智能可以看作是对传统生产工具的替代, 人工智能技术为提高获取信息数据能力、提高人类时间利用效率、拓展人类社会空间等方面作了积极贡献, 推进人类不断建构新的需求, 只不过人工智能技术更为强大和复杂。如果传统机器是对人类体力劳动的延伸, 那么, 人工智能这种新兴科技便是对人类体力劳动和脑力劳动的创新突破。至此, 可以说人工智能与传统机器一样都是人类发挥自身力量的手段, 服务于人类自身发展的现实需要。

总而言之, 人工智能是物理性的智能人工物, 在生产过程和决策过程中执行着人的意志和目标。人工智能虽然在某种程度上超出了人们的预想的范围, 逐渐具备类人智慧与拟人特征, 但人工智能无法模拟人类的右脑和边缘系统(非逻辑思维), 无法脱离人类智能独立思考。“人工智能所面临的难以真正具有人的心灵表征事物的状态属性的意向性、只识程序上的语法而无法通晓人类的语言意义本身等发展困境, 使它与人工智能依然会存有显著的差别”^[9]。目前, 人工智能依赖于人类的科技进步、受限于科技发展的所达程度。虽然ChatGPT在自我学习、自我进化和自我创造等综合能力方面取得了突破性进展, 甚至在各种专业和学术基准测试中已达到人类水平, 但它仍旧是以大型语言模型GPT-3为基础技术, 依旧是人工智能现实化的表征方式。

2. 人工智能技术应用的风险类型

自人工智能诞生以来, 社会各界对其关注度

呈现出持续高涨的态势。ChatGPT的横空出世不仅可以学习和理解人类的语言和思维方式,具备类人思维与人类进行沟通,而且可以自动生成各种创造性的内容,是继Stable Diffusion图像内容生成后的又一里程碑式模型,更符合人类交往的情感需求。ChatGPT以崭新的方式丰富了人们的物质生活体验表明:“人工智能具有广泛的学习能力并在大多数领域达到或超过普通人类的能力”^[10]。但ChatGPT随即也带来了一些不良影响,其在歧视、数据泄露、网络安全、知识产权等方面都面临诸多潜在隐患。人工智能技术应用风险论从本质上主要分为四种类型。

2.1 技术风险

技术风险指在风险社会中“由于技术进步而引发的人类社会发展的不确定性,以及这种不确定性对人类生存、发展的不利影响”^[11]。主要表现为人工智能技术本身存在的缺陷或人类错误地使用导致的风险挑战。人工智能技术“主要是基于数据分析和模型建构来进行深度学习,并通过建立数据变量之间的关系而做出自我决策”^[12],这种风险意味着人工智能技术在与人类的关系中处于被动从属地位,它并不具有自主“威胁”人类的意图。

人工智能技术在应用过程中产生的技术风险主要表现为三种情状:一是由于技术本身存在的缺陷而引起的技术风险,主要表现为算法结构缺陷、非鲁棒性频繁出现等,导致人工智能技术以错误的数据、模型、代码为基础进行运作,最终必然会给人类自身造成威胁。二是由于人工智能技术囿于算法黑箱、算法歧视和算法偏见的存在,使算法具有不透明性和运行的自主性,无法对技术内部的运行规律和因果逻辑机制提供有效的解释,人们难以理解和解释技术的生成机理,从而引发一系列未知的隐私权、知情权等被侵犯的风险。数据隐私泄露问题已经不再局限于传统的公众领域,“数据与算法所构建的‘超级全景

监狱’致使处其中的人们已毫无隐私可言”^[13]。

三是由于人工智能技术的不合理应用产生的风险挑战,直接或间接地伤害着人类生存发展。比如部分不法分子利用人工智能技术非法采集个人数据隐私,导致“智力拜物教”的出现。由于人工智能技术的系统运作依赖于海量信息,但在数据流转过程以及人工智能技术模型都有泄露隐私数据的可能性。人类被人工智能技术裹挟至智能化系统中,个人的基本信息、浏览记录、兴趣爱好极易被强大的算法进行实时追踪,严重威胁着个人对自己隐私的掌控权。

2.2 法律风险

法律风险主要是指人工智能技术的突破性发展以及广泛服务于社会公众的同时,现有法律体系难以对其进行有效规制而引发的法律失效风险。探讨人工智能技术法律风险问题,需要以人工智能行为的可解释性为出发点,进而对人工智能技术的法律风险进行合理的预估和规制。

现有人工智能技术应用法律规范问题具体包括:一是人工智能技术引发的侵权责任风险,难以确认其是否具有独立的法律人格问题。虽然Open AI公司针对ChatGPT生成内容的商业性问题制定了法律法规加以约束,但“不同用户同时输入相同问题(甚至差异性问题的)得出相同内容的利益归属及相关纠纷解决路径未臻明确”^[14]。人工智能技术发展使其技术产品逐渐具备类人思维,一旦发生人工智能技术侵害他人权益的问题,现有民法是否可以将其作为独立的民事主体,究竟谁该承担侵权责任?二是人工智能产品超出原先设定程序而造成的责任归属问题,挑战现有社会伦理秩序。人工智能技术决策引发的法律问题主要指人工智能技术本身并没有触碰法律的底线,但其决策结果在使用过程中造成了违法行为。同时,现有法律体系并未直接承认人工智能技术的法律主体地位。因而,人工智能技术潜在的道德

伦理问题在实际应用中具有进一步扩大的风险。三是人工智能产品在实际运用场景中面临归责体系重构的难题。如特斯拉汽车自动驾驶系统发生交通事故造成人员伤亡事件, 究竟是汽车所有人承担责任还是自动驾驶系统的设计者承担责任? 可见, 对于事故责任的重新认定不再以过错为基础, 而是将责任指向了自动驾驶系统的开发者和制造者, 这预示着旧有的道德归责体系将面临重构的困境和风险, 挑战现有法律体系。

2.3 伦理风险

伦理风险是指在人机融合智能技术的发展过程中, 人们在技术发展和伦理现实的冲突中违背了传统道德习俗和社会舆论, 从而引发的道德危机和伦理困境。主要体现为: 首先, 人工智能技术的算法歧视。算法歧视再现了人类社会歧视。算法本质上不具有歧视性, 但技术设计者的道德预置不可避免地造成算法系统的偏见。算法歧视主要包括算法控制者输入数据歧视、关联歧视和“大数据杀熟”三种类型, 损害了公民基本的平等权利、具体的隐私权、公平交易权等权益、市场公平竞争秩序。其次, 人工智能技术的算法权力。算法权力将挑战传统权利格局, 科技精英逐渐独掌知识权威局面, 形成横向权利模式。算法权力可以学习和分析现有的信息数据, 将自身的价值转换为“准公权力”, 进而达到操纵社会的目的。一方面, 公权力可以凭借算法权力实现对公民个人信息的大量采集, 并且实现对某个人的全方位跟踪和掌控; 另一方面, 企业组织可以通过自身掌握的算法技术优势, 在社会公共生活领域掌握更多的话语权, 进一步巩固自己的地位。最后, 人工智能技术加剧了数字鸿沟和决策偏斜。人工智能在采集数据时过度依赖训练数据, 如果数据存在偏见和歧视, 它会有意无意地排斥一部分弱势群体, 导致数字鸿沟和不公平智能决策的出现, 这将更加不利于贫困人群、低认

知人群在社会的生存, 加剧社会各阶级的分化。同时, 人工智能引发“信息茧房”的现象时常发生。人工智能不仅可以让用户自行进行虚拟交往, 而且利用强大的算法能力对已有数据进行精准推送, 用户逐渐聚合为信息群体。长此以往, 用户将自身牢牢地管制于信息茧房内, 造成用户与其他个体、社会群体间的疏远, 渐渐丧失社会交往能力。

2.4 社会风险

社会风险指人工智能技术发生重大变革导致的个人、群体乃至整个社会对现有生产生活方式的“不适应”现象, 将挑战人类主体性地位。新技术革命引发的社会性失业问题是适应性威胁的典型表现形式。随着人类与机器之间的界限日益模糊以及人工智能核心关键技术的迭代升级, 智能设备主体性呈现出日益增强的趋势, 不可避免地冲击和挑战人的主体地位。人工智能技术充分利用现有的社会资源, 极大地提高了人类的生产和生活效率, 缓解了人类的劳动压力。但人工智能技术也引发了诸多的社会风险, 主要表现为: 一是人工智能技术应用对劳动者工作的侵蚀, 去劳动化趋势愈加明显。人工智能技术不仅可以辅助人类完成部分繁重、单调、重复的体力劳动和脑力劳动, 而且可以独自承担有毒有害的岗位职责, 从而引发不同行业、岗位和工种劳动群体之间的矛盾冲突。人类为了满足生存的基本需要不得不调整劳动方式, 处于社会边缘地位的低端岗位和智能产业所需的专业人员将得到绝对工作优势, 那些处于中等地位的人员将面临失业的风险。另外, 人工智能使劳动者从生产劳动中剥离出来, 形成新的非生产性劳动就业形式。

“数字零工经济”是这种劳动形式的典型代表, 它看似公平自由, 实质上是资本对劳动者的劳动力的购买, 背后隐藏的仍然是资本家对劳动者的绝对剥削。二是人工智能技术将不可

避免地形成行业垄断，导致企业间的恶性竞争，妨碍市场机制的正常运行。在市场中，部分利益既得企业凭借掌握的关键核心技术优势获得操纵和垄断地位，通过限制竞争、强买强卖、限价限产等行为，肆意打压中小企业或歧视合作商家，从而达到谋取巨额利润的目的，严重破坏和阻碍了市场竞争秩序，损害了相关消费者和经营者的利益。

目前，就人工智能在当前以及可预见的领域而言，人工智能技术代表了生产力发展的最新形态，为实现人的本质复归创造了良好的条件，但也给人类社会带来了诸多风险挑战。随着人工智能技术不断更新和突破，人工智能逐渐颠覆人类传统的劳动概念和形式，呈现出否定人类劳动的趋势。人工智能威胁论问题从源头上看是由人工智能技术的实际应用造成，但必须明确，人工智能目前仍然处于弱人工智能阶段，只能解决目标领域的单一问题，它背后隐藏的依旧是人与人之间的现实关系。

3. 人工智能技术应用风险的规避路径

马克思就第一次工业革命推动生产力水平的快速提高指出：“自然科学却通过工业日益在实践上进入人的生活，改造人的生活，并为人的解放做准备”^[15]。人工智能技术为推动经济社会形态的超越提供了新的契机，极大地促进了生产力发展和社会关系解放。未来数字社会将呈现出走向信息化和智能化相结合的趋势，亟须有效的手段对人工智能技术应用风险进行防范和规制。因此，在全面建设社会主义现代化强国的历史节点上，我们不能无的放矢、不加以变通地寻找一种永恒的应对策略，而是要对人工智能技术应用保持谨慎态度，立足我国人工智能技术发展实际，构建具有中国特色的人工智能技术应用风险治理综合方案。

3.1 加强技术伦理规范和科学家的人文精神，提高人工智能技术模型的可解释性和可追溯性

当前社会中的风险种类繁杂且彼此互相联系，对人工智能技术应用风险要求必须对各类风险建立系统的认识。事实上，人工智能技术本身是中立的不具有道德属性，但研发者的“道德预置”决定了人工智能应用的道德水准和价值取向不可避免地会带来一些伦理挑战和风险，需要进一步实现人工智能技术变迁与社会伦理价值体系之间的相互协调。

首先，增强人工智能技术工作者的人文精神，提高人工智能系统的公平性。科研工作者的职责主要是致力于为人类和其生存环境谋取最大化的福利，需要自觉接受伦理体系的制约。相关部门应尽快出台技术工作者伦理行为规范，“制定新兴技术伦理规范和标准，为新兴技术研发和使用主体提供行为向导。”^[16]在研发阶段应尽量降低加入个人主观因素，使算法的结果更具客观性和规范性；在数据层面利用完整异构的数据集将数据固有歧视和偏见最小化，同时对数据集进行周期性检查确保数据的高质量性；在技术应用层面需要引入公平的算法减轻或者消除决策偏差和潜在的歧视。

其次，健全人工智能技术应用风险研判机制，提高对人工智能技术应用引发的社会风险的掌握度。一方面，政府应该及时更新风险储备。当前，人工智能技术实际应用中引发的社会性失业、数字鸿沟、道德失调、算法歧视等都是之前社会中未出现的新增风险，需要社会各方面及时更新风险观念并且做出应对之策，有效防范技术应用新风险对人类社会带来的危害。另一方面，政府应该制定更加详细的人工智能技术风险预案，对人工智能技术发展趋势进行科学预测和跟踪研究，在获得人工智能技术发展的社会红利之外加强对可预见社会风险的治理。

最后，加强人工智能技术伦理规范，自觉抵制算法黑箱。人工智能技术本身负载着一定的社会价值，技术的价值负载实际上是“内在于技术的独特的价值取向与内化于技术中的社会文化价

值取向和权力利益格局互动整合的结果”^[17]。针对人工智能算法存在的固有缺陷, 需要提高人工智能算法决策系统的可解释性和可追溯性, 减少不同主体之间的信息不对等现象。算法透明度问题是人工智能技术工作者与人工智能技术使用者——社会公众之间的信息不对等, 需要提高人工智能技术算法黑箱运行机制的透明度, 使算法的运行规律和因果逻辑更易被公众理解和监督。算法所有权问题是人工智能技术企业和社会公众之间对信息数据掌握的差异性。数据隐私特有的繁琐性、复杂性和模糊性, 人工智能技术利用数据所有权将个人隐私置于被采集的风险中, 社会公众在现实中也表现得十分被动, 难以真正掌握个人隐私。为此, 需要政府层面建立更为细致的用户隐私条款和多元化选择, 确保个人在使用人工智能技术中对自身隐私的所有权。算法公平性是由人工智能技术算法透明度和所有权问题导致的现实社会问题, 主要表现为算法歧视和算法偏见, 需要尽量降低算法运行受数据、模型等因素的影响, 确保算法系统的正常运作, 用技术手段为算法进行去偏见化处理, 使算法结果更具客观性和规范性。

3.2 推动人工智能技术运用治理的国际合作, 构建人工智能技术伦理与法律标准化治理体系建设

从整体上看, 全球人工智能治理工作取得初步成效, 正从宏观的伦理道德标准向微观的政策法律等实践措施推进。人工智能技术伦理风险评估和政策规避路径的研究已然成为全球治理共识, 各方致力构筑具有整体性和具体性的人工智能配套规范标准。截至目前, 世界各国相继出台人工智能相关法律法规、政策文件和公告, 为我国人工智能治理工作提供了良好典范。2022年, 英国政府发布《国家人工智能战略》中强调: “人工智能治理和监管应该紧跟其快速变化的需求, 英国需要在这方面有所作为以达到持续创

新、保护国家和公民安全的目的”^[18]。2023年, 美国国家标准与技术研究院(NIST)正式公布《人工智能风险管理框架》(AI RMF 1.0), “旨在指导组织机构在开发和部署人工智能系统时降低安全风险, 避免产生偏见和其他负面后果, 提高人工智能可信度, 保护公民的公平自由权利”^[19]。同时, 英国国防部国防科学技术实验室计划建立人工智能研究团队, 提高其在人工智能和数据科学领域的能力。可见, 各个国家和地区加大了在人工智能领域的国际合作和结盟, 力图在人工智能技术发展和运用过程中抢占先机。

我国对人工智能技术应用治理工作始终将重点聚焦于人工智能的开发与应用阶段, 致力于实现二者的双向平衡治理。当前, 我国人工智能技术治理的伦理与法律协同治理机制整体框架已初步完成, 逐步构建起兼顾安全保障与发展效率的治理模式。2017年, 我国政府为抓住人工智能技术发展的重大战略机遇, 构筑起我国人工智能发展先发优势, 国务院印发《新一代人工智能发展规划》, 对人工智能技术发展安全进行宏观部署; 2019年, 国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》, 提出了人工智能技术治理的具体框架和行动纲领。但整体而言, 我国人工智能技术应用风险的治理仍然处于初步发展阶段, 具体的风险治理政策相对短缺和僵化, 还需要进一步合理地拓展人工智能技术的政策规划。事实上, 人工智能技术风险治理的伦理规则与法律规制要求具备增进人类社会福祉的战略目标和价值原则, 兼顾政府、企业、科研机构等多主体以及计算机科学、法学、社会学等多学科的多元协同, 按照可解释性、安全性、人类可控性的标准和应用审核、事后追责、权利保障的方法协调各方群体利益目标达成, 稳定推动协同治理系统的可持续性发展。同时, 人工智能的发展仅仅依靠法律法规的规制作用难以达到治理的理想效果。

我国应当充分发挥政府监管的主导作用，建立完备的监管体系和公开透明的监管组织，协调群众、教育机构、其他行业等各方力量，共同加强对人工智能技术应用的监管，最大限度规避人工智能所引发的现实困境。

3.3 以社会合作共治为框架，推进人工智能技术风险协同治理

习近平总书记在党的二十大报告中明确提出：“健全共建共治共享的社会治理制度，提升社会治理效能”^[20]。由于人工智能技术创新的直接动力来自于市场需要，其背后汇聚着不同的价值理念甚至政治动机，因此针对人工智能技术应用风险的治理，需要采取社会合作共治的复合治理模式。构建自上而下的治理制度体系，从政府部门、社会、企业和公众层面实现各环节配套衔接，形成政府监管、社会协同、企业合作、公众参与的社会治理格局，发挥多元主体的敏捷性和灵活性，提升人工智能技术应用风险的治理效能。

国家层面，应发挥政府主导效能。政府在人工智能技术应用风险治理的过程中应该扮演如下角色：一是人工智能技术负责任发展政策的制定者，确保技术应用的可持续性。由于人工智能技术应用本身存在的技术和社会双重不确定性风险，政府要将人工智能技术负责任发展的价值诉求转变为切实可行的政策举措。政府应制定和完善人工智能技术应用问题相关的法律法规，使人工智能技术在法律的规范下良性和向善发展，进而在实际应用环节减少技术风险。二是治理议程的设计者。治理议程指政府对议题的重要性与否进行先后排序。针对技术研发和应用企业而言，由于他们扮演既得利益者角色，因而不会提出与算法歧视相关的治理议程，但事实上，算法歧视对公众和社会的公平正义已经造成严重威胁。因此，在人工智能技术应用风险的治理中，政府

应该扮演治理议程的设计者角色。三是法律法规的监督者。一方面，在人工智能技术应用风险治理中，政府应成立人工智能专业委员会，并且明确各专业委员会的地位和职能。由专业委员为行政执法人员提供技术方面的决策咨询服务，以此弥补一般人员的专业知识欠缺问题。另一方面，政府应承担起提高人工智能技术运作的透明度任务。由于人工智能技术的内部运行系统和工作原理对普通用户而言可以是不可知的，但对于技术的设计者和研发者而言，必须取缔算法黑箱的存在，提高算法的可解释性和透明度。

社会层面，应发挥社会引导效能。超级数字平台本身掌握了先进核心技术和海量数据信息，新闻媒体逐渐成为数字化社会监督和宣传的关键工具，然而在人工智能技术应用风险领域它们扮演的角色却是微乎其微，发挥社会媒体治理的能力更是微不足道。社会媒体可以为人工智能技术应用风险的治理提供信息交流和共享的崭新平台。一方面，社会媒体对人工智能技术应用过程中产生的风险信息进行系统分析和研判，更好地保护公众的切身利益，不断地引导公众参与技术应用风险的治理中；另一方面，社会媒体嵌入人工智能技术应用风险治理，使社会媒体成为治理的重要载体。社会媒体服务于社会公众，为公众搭建了平等的话语平台，充分发挥新型主流媒体和全媒体传播体系的引导作用。然而，社交媒体在人工智能技术应用风险治理的效能发挥的过程中是一把“双刃剑”，其治理效能也会大幅削弱。为了充分发挥社会媒体对人工智能技术应用风险治理的监督 and 宣传效能，需要政府与公众建立新媒体沟通渠道，鼓励社会媒体积极参与人工智能技术应用风险治理中。在实际运行中，提升社会公众的媒介素养，建构网络空间中的公众信任体系；培养社交媒体的责任意识，构建公开透明的社交媒体传播机制。

公众层面，应发挥公众补充效能。社会公众

是人工智能技术应用的直接参与者、亲历者和实践者。随着智能时代的不断推进,在信息化技术的裹挟下公众逐渐参与到人工智能技术应用治理的实践中。然而在实际治理过程中,公众对人工智能技术应用的治理参与度低下。为此,需要政府大力宣传和引导社会公众参与到人工智能技术应用风险的治理中,营造公众积极参与风险治理的社会氛围。一方面,切实增强公众参与技术风险治理意识,引导公众从被动参与转向自觉参与技术风险治理。发挥政府和社会的职能优势和信息优势,引导与人工智能技术相关的科学家积极参与到科普活动中,使公众更加详细地了解与人工智能技术应用风险治理的有关举措和信息资源,切实保障公众的知情权和监督权。对参与技术风险治理成绩突出的企业和个人进行及时宣传报道,逐步提高公众参与风险治理的能力和水平。另一方面,健全公众参与风险治理的参与渠道和政策机制。公众积极参加政府和社会组织的与人工智能技术应用风险治理主体相关座谈会,政府和企业对公众提出的治理方法、策略和意见积极开展讨论和试点,有效汇集人工智能技术应用风险治理的公众力量。同时,公众可以对人工智能技术应用风险治理举措进行积极的反馈,促使举措部署者及时改进,促进治理举措与道德价值相匹配。

4. 结语

总而言之,人工智能技术作为全球前沿技术创新领域的一项崭新技术,并未完全发展成熟。人工智能技术在带来其他产业领域的颠覆性发展并促进人类实现诸多领域便利的同时,也引发了一系列难以应对的风险挑战。因此,面对人工智能技术发展带来的机遇和挑战,盲目限制人工智能超越图灵奇点,甚至放弃人工智能发展的做法,无异于因噎废食。不仅要充分释放人工智能技术发展的红利促进人类社会综合性变革,而且

需要充分发挥社会治理的作用。针对人工智能技术存在的技术缺陷,需要加强技术伦理规范和科学家的人文精神,提高人工智能技术模型的可解释性和可追溯性;针对人工智能技术存在的治理缺陷,需要以社会合作共治为框架,协同政府部门、社会、企业和公众层面实现各环节配套衔接,形成政府监管、社会协同、企业合作、公众参与的社会治理格局;针对人工智能技术存在的治理难题和差异性,需要推动人工智能技术运用的国际合作,推动人工智能技术伦理与法律标准化治理体系建设。

责任编辑:李琦 校对:李琦 梁思琪

参考文献

- [1] 樊博.ChatGPT的风险初识及治理对策[J].学海,2023(02):58-63.
- [2] Cath Corinne. Governing artificial intelligence: ethical, legal and technical opportunities and challenges[J]. *Philosophical Transactions*,2018,376(2133):1-8.
- [3] Bullock J B. Artificial intelligence, discretion, and bureaucracy[J]. *The American Review of Public Administration*, 2019,49(07):751-761.
- [4] 张欣.生成式人工智能的算法治理挑战与治理型监管[J].现代法学,2023,45(03):108-123.
- [5] 张辉,曾雄,刘鹏.中国式治理现代化视域下的人工智能治理[J].中国科学基金:1-8.
- [6] 马克思恩格斯文集(第1卷)[M].北京:人民出版社,2009:501.
- [7] 高伟.人工智能与人类解放的未来——基于马克思的科技哲学观[J].人民论坛,2020(10):122-123.
- [8] 马克思恩格斯全集(第31卷)[M].北京:人民出版社,1998:102.
- [9] 程萌.人工智能奇点论的历史唯物主义批判——从"物种奇点论"到"社会奇点论"[J].学习与实践. 2023(03):121-129.

- [10] 何哲,曾润喜,秦维,等.ChatGPT等新一代人工智能技术的社会影响及其治理[J].电子政务,2023(04):2-24.
- [11] 毛明芳.应对现代技术风险的伦理重构[J].自然辩证法研究,2009,25(12):55-60.
- [12] 唐要家,唐春晖.基于风险的人工智能监管治理[J].社会科学辑刊,2022(1): 114-124+209.
- [13] 冯永刚,席宇晴.人工智能的伦理风险及其规制[J].河北学刊,2023,43(03):60-68.
- [14] 邓建鹏,朱恽成.ChatGPT模型的法律风险及应对之策[J].新疆师范大学学报(哲学社会科学版),2023,44(05):91-101.
- [15] 马克思恩格斯文集(第1卷)[M].北京:人民出版社,2002:301.
- [16] 安慧影,黄朝峰,李阳.新兴技术伦理风险协同治理研究[J].科技进步与对策,2023(3):1-10.
- [17] 刘大椿.科技与伦理何以内在统一[J].山西师大学报(社会科学版),2023(1):53-60.
- [18] 中国科学院.英国政府发布《国家人工智能战略》[J].网络安全和信息化动态,2021(10).
- [19] 中国科学院.美国国家标准与技术研究院发布《人工智能风险管理框架》[J].网络安全和信息化动态,2023(3).
- [20] 习近平.高举中国特色社会主义伟大旗帜为全面建设社会主义现代化国家而团结奋斗——在中国共产党第二十次全国代表大会上的报告[M].北京:人民出版社,2022.

Ethical risks and avoidance ideas in the application of artificial intelligence technology

Yang Li, Liu Wenwen^{*}

(Lanzhou University of Technology School of Marxism, Lanzhou, Gansu 730050, China)

Abstract: With the rapid development and wide application of artificial intelligence technology, it has become the consensus of global governance to strengthen the research on the ethical risk prediction and policy avoidance path of artificial intelligence technology. Therefore, the nature of artificial intelligence technology should re-examined, and the technical, ethical, social and legal risks brought by the application of artificial intelligence technology should be faced up. In fact, the hidden danger avoidance of artificial intelligence technology embedded in social practical application is a complex system engineering, which should be committed to building a comprehensive governance framework. In short, risk awareness should be improved, technical ethics and the humanistic spirit of scientists should be formulated and fostered, and the interpretability and traceability of AI technology models should be enhanced. With social cooperation and co-governance as the framework, propel the collaborative governance of artificial intelligence technology risks, impel international cooperation in the application of artificial intelligence technology, and accelerate the construction of a standardized governance system for artificial intelligence technology ethics and laws.

Keywords: artificial intelligence technology; ethical risks; circumvention path