人工智能发展的风险与治理

石琳娜

四川省科学技术发展战略研究院,四川 成都 610000

摘要:目前,以 ChatGPT 为代表的人工智能技术引发了广泛关注。人工智能发展带来了一系列风险挑战:主体性危机引发人类自身主体性质疑,丧失自主判断力,带来现实人际孤岛;数据缺失造成结论误导,虚假数据造成信任危机,数据泄露和滥用带来安全隐患;算法的可解释性不足带来“黑箱”问题,算法的鲁棒性不足造成算法失灵,“回音室效应”带来偏见与分化;社会风险包括技术的颠覆性变革带来国际分工重塑,意识形态的舆论引导性风险,行业科技巨头垄断趋势加剧;等等。人工智能治理在中国、美国、欧盟得到了很好的实践。敏捷治理、韧性治理、负责任创新和治理,是最新的人工智能治理的三大理念。对于人工智能的发展和治理,不仅要充分释放“制”的优势,更要做好“治”的文章,充分发挥各治理主体的作用,需要政府发挥统领监督引导作用,整合企业、科研机构、高校的各种资源,实现政府、企业、高校、科研机构的协同发力。

关键词:人工智能;治理;主体性危机;算法

中图分类号:F224.32 **文献标识码**:A DOI:10.12186/2023.02.008

文章编号:2096 - 9864(2023)02 - 0056 - 07

当今世界,全球重大前沿技术和颠覆性创新快速突破,新一轮科技革命和产业变革带来重大机遇,以人工智能、大数据、云计算、区块链为代表的新一代信息技术飞速发展、加快渗透,对人类的生产、生活方式产生了深刻影响。一方面,以人工智能为代表的新兴技术与经济社会深度融合,催生了智慧城市、智慧农业、智慧能源、智慧交通等新领域新赛道,成为我国构建现代化产业体系、提高经济综合竞争力的重要战略引擎;另一方面,人工智能等新兴技术在加速迭代升级的过程中,也给人类社会的法律隐私、伦理道德、社会治理、国际关系准则等方面带来一系列问题和挑战。

关于人工智能的研究已经成为全球性、全局性的问题,国内外学者从不同角度对人工智能展开了研究。W. J. Rapaport[1] 认为,人工智能属于计算机科学的一个分支,是一门通过研究算法来解决问题和完成任务的学科;M. Maas[2] 认为,人工智能的内涵包括了众多内容,目前是一个复杂而又混乱的概念;庞祯敬等[3] 认为,人工智能催生出的诸多新产业、新服务、新业态和新模式已成为经济社会数字化变革的核心动力;闫德利[4] 认为,人工智能的发展将引起就业结构的巨大变化,单一特定领域的重复性工作将被人工智能大量取代;石琳娜等[5] 认为,以人工智能、云计算、物联网等为代表的数

收稿日期:2023 - 03 - 16

基金项目:四川省科技厅软科学研究项目(2023JDR0001、2023JDR0145);河南省哲学社会科学规划项目(2021BZZ013)

作者简介:石琳娜(1987—),女,河南省平顶山市人,四川省科学技术发展战略研究院副研究员,中组部“西部之光”访问学者,主要研究方向:数字经济与科技创新治理。

字经济重要性日益突出,并与实体经济加速融合;鲁传颖等[6]认为,人工智能风险存在于各个领域,对国际、国家、社会、个人等都产生了广泛影响;洪永淼等[7]认为,以 ChatGPT 为代表的前沿人工智能目前仍没有人类的意识或理解能力,只有预测能力;T. Liu[8]认为,增强人工智能的可解释性成为热点问题,但仍处在初期研究阶段,有多种研究路径可供持续探索;梁正等[9]认为,世界诸国对于什么是人工智能治理、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。

综上所述,目前国内外关于人工智能治理的研究还处于起步和探索阶段,理论研究的步伐明显落后于实践,在我国新发展阶段和经济社会转型升级的背景下,探讨人工智能治理相关问题具有重要意义。

一、人工智能发展带来的风险挑战

目前,以 ChatGPT 为代表的人工智能技术引发广泛关注,人工智能等新兴技术在加速迭代升级的过程中,也引发了一系列风险和问题。

1. 主体性危机

(1)引发人类自身主体性质疑

当前人工智能系统在某些方面已经可以与人类相媲美,甚至超越人类,人工智能的普及会削弱人在社会中的主体地位。例如,人工智能系统已经开始取代人类部分工作,这可能会使人们开始怀疑自己的职业价值和发展前景。未来,人类有可能逐渐难以参与智能生产的过程,丧失在实践活动中的主体地位。

(2)失去自主判断力

人工智能所具有的高效信息收集、分析、推理、决策能力是人类所难以企及的,但当人们过度依赖人工智能时,他们可能会失去自己的判断能力。另外,人工智能系统的设计和训练可能存在偏见和局限性,如果人们长期依赖它,可能会失去对这些错误的察觉和纠正,导致人们的自主判断和决策能力被进一步削弱。

(3)带来现实人际孤岛

人工智能技术实现了虚拟社交的扩大化。当今社会,网络社交平台层出不穷,打破了时空和地理的限制,让全球的用户可以同时在线使用并分享和交流相关信息。这种虚拟社交的扩大会促使用户快速获得自己喜欢或认同的内容,筛选与自己志趣相投的群体、与自己有共同语言的虚拟人物。通过算法设计出来的虚拟人物能充分满足用户需求,比真实人类更加善解人意,这样会极大地降低现实世界人际交往的欲望和动力。同时,对人工智能产品的大量使用会分散真实世界中人类群体建立亲密关系的精力,拉大身边的人与人之间交往的鸿沟,人际孤岛的问题也逐渐凸显。

2. 数据问题

(1)数据缺失造成的结论误导

人工智能进行自动化决策时,如果数据不充分、不达标,会导致预测结果不准确、不可靠。例如,医疗领域的人工智能模型在训练数据中如果缺少某些疾病的信息,就会出现漏诊或误诊从而给患者带来严重的健康风险;在金融领域,如果缺少某些欺诈行为的信息,就会产生误判从而导致金融损失和信用危机等。

(2)虚假数据造成的信任危机

人工智能算法具有极高的仿真度,使得日常活动中的防伪鉴定面临困难,对社会各个层面都构成潜在的威胁。如果在算法中混入虚假的数据还会对算法形成欺骗,在智能决策结论中得出错误的结果。当人工智能模型故意模拟或使用虚假数据时,可能会导致预测结果的低准确性,从而降低用户对模型的信任。

(3)数据泄露和滥用带来的安全隐患

数据的采集和使用若管理不到位会给个人、社会和国家带来安全隐患。例如,企业在加

工、使用数据的过程中出现泄露和滥用,不仅会威胁到用户的个人隐私,还会对企业乃至社会形成巨大风险。此外,如果国家机密数据被泄露,就可能会被用于进行间谍活动、网络攻击等,对国家安全造成威胁。

3. 算法的固有缺陷

(1)算法的可解释性不足带来"黑箱"问题

普通用户无法理解程序的复杂运算过程,可解释性不足让人们不能理解算法的决策机理,也难以预测算法的行为,"黑箱"问题由此出现。深度学习算法的一个显著特点是在训练过程中自动提取特征,但这一过程目前尚不可控,算法可能会选择错误的特征。算法的透明性和可解释性不足会限制技术的传播和发展,成为算法利用方面的难题。

(2)算法的鲁棒性不足造成算法失灵

算法的鲁棒性是指算法在处理输入数据时,能够保持稳定性和正确性的能力。一个具有鲁棒性的算法,可以正确地处理各种类型的数据,包括缺失数据、错误数据、离群值和噪声等。当算法的鲁棒性不足时,它可能会输出错误的结果,或者出现错误时无法正常运行,导致算法失灵。

(3)回音室效应带来偏见与分化

深度学习算法能够挖掘训练数据集中不同因素的相关性,拟合其分布特性,数据集中的偏见与歧视会被引入到训练出的模型之中。在网络信息化时代,各种数据和信息可能会被分类和筛选,并进行标签固化,在算法推荐系统的不利影响下很有可能产生"回音室效应",导致人们长期获得自身画像的单一数据和信息,致使人们沉浸在自己偏好的信息世界中,可能造成严重的公众意识形态分化、网络群体极化现象[10]。

4. 社会风险

(1)技术的颠覆性变革带来国际分工重塑

人工智能等新兴技术的变革发展,也带来了科技保护主义的兴起。美国在算力、算法、芯片等关键领域都有较大优势,基于自身的先行优势,美国有能力构建一个新的、更固化的中心 - 外围结构的世界分工格局。近年来,美国依靠自身的科技和经济实力,以及自身的世界霸主地位,正在强化单边科技保护主义,努力与中国以及其他国家脱钩,以保持自身的领先地位。其他发展中国家与美国在人工智能领域,特别是在通用模型的科技研发方面差距巨大。科技力量悬殊加剧了广大科技弱势国家对发达国家的技术依赖,科技弱势国家被迫选边站队,科技政治极化风险激增。作为最大的发展中国家,中国在人工智能的模型迭代创新方面拥有较大上升空间,仍需要不断努力。

(2)意识形态的舆论引导性风险

以 ChatGPT 为代表的人工智能大模型高度拟人化和便利性将强化社会公众对它的信任和依赖,同时增强它对重大概念和事件的解释权,与政府官方的解释权形成竞争,其输出结果的真实性持续影响公众的判断和选择,潜移默化之中塑造和改变公众的观念和舆论判断。拥有技术优势的国家有可能利用他国公众的从众心理,利用人工智能大模型输出意识形态偏好,强化西方价值观,隐蔽进行意识形态渗透和制度干扰,侵蚀技术弱势国家的政权管理能力。同时,国际互联网平台充斥着的不计其数的智能水军,也给技术弱势国家带来了巨大的监管压力,增加了社会不稳定因素。

(3)行业科技巨头垄断趋势加剧

人工智能作为前沿新兴技术,在理论和实践中拥有广阔的发展空间。但是,在实践应用和落地发展中,通用大模型并不"通用",该领域的研发和创新只有科技巨头才能涉足。目前,用于存储数据模型的显卡至少价值 200 万元,模型迭代需要长期、大量的资本、人力投入,其需要的算力和数据储备只有科技巨头企业才

能满足,得体的交互应对背后是聚沙成塔式的细节更新,高昂的训练和能源消耗成本也不是普通中小企业能提供的。科技巨头企业高额投入换来的技术不断更新进步,同时带来的经济回报和影响力也是巨大的,长此以往,这种趋势带来的科技垄断将更加明显。

二、人工智能治理的全球实践

自 2016 年起,已有 40 多个国家和地区将推动人工智能发展上升到国家战略高度,加快推动人工智能的创新发展。ChatGPT 的出现反映了人工智能发展的新趋势,即 AI 正在从感知智能向认知智能快速发展。作为人工智能领域发展最先进的国家,美国拥有一系列具备充足技术和资金资源的公司和实验室,如谷歌、微软、OpenAI、亚马逊、Meta 等科技巨头,都在通用大模型上进行了大规模布局与投资。继 OpenAI 发布 ChatGPT 大模型之后,中国的科技公司也陆续加入了大模型开发和应用的队伍,并努力将之前的发展重点(单一模型对应单一任务的专用大模型)调整为通用大模型,如百度开发的“文心一言”,阿里开发的“通义 M6”,腾讯开发的“混元”,华为开发的“盘古”,中国科学院自动化研究所开发的“紫东太初”等。人工智能的快速发展引起了全球的广泛关注,与此同时,开展人工智能治理,发展负责任和可信的人工智能也正在成为全球共识。中国、美国、欧盟作为三大经济体是最具有代表性的典型样本[9],近几年,已经相继出台了人工智能治理的相关法规,做到了有法可依。

1. 中国实践

为抢抓人工智能发展的重大战略机遇,构筑我国人工智能发展的先发优势,加快建设创新型国家和世界科技强国,我国在大力推动人工智能发展的同时,也出台了一系列政策法规,指导人工智能的健康发展。国务院在《新一代人工智能发展规划》中提出,“建立人工智能法律法规、伦理规范和政策体系,形成人工智能安全评估和管控能力”。该规划强调防范人工智能风险,引导人工智能向有益于人类的方向发展,需要及时就人工智能出现的风险及其防范机制进行研究。国家标准化管理委员会发布的中国《人工智能标准化白皮书 2018》,首次从政府层面提出了中国关于人工智能伦理的人类利益原则和责任原则。同时,中国也通过立法的形式进一步规范以人工智能为代表的新兴技术的发展,先后出台了《数据安全法》《个人信息保护法》《网络安全法》,我国人工智能治理逐渐进入法制化轨道。我国各部委也持续跟进相关政策文件的补充落实,如由国家网信办发布的以算法为治理对象的政策文件《互联网信息服务算法推荐管理规定》,对企业使用算法时需要遵守的规定、违法之后的处罚作了相应规定。由工信部发布的《“十四五”信息化和工业化深度融合发展规划》,对智能产品在行业的应用推广进行了系统性的部署,对其他部署也做了相应的行为规定[11]。

2. 美国实践

美国作为科技实力最强的国家,其人工智能技术发展领先全球。美国先后出台了三个扶持人工智能发展的政策文件《国家人工智能研发战略规划》《为人工智能的未来做好准备》《人工智能、自动化与经济报告》,同时渐进式推进人工智能治理的立法实践。在美国,各州率先出台了人工智能治理相关的法规,但是在联邦层面的立法进展始终较为缓慢。在治理方式上,美国政府问责局发布了一个人工智能问责框架,帮助联邦机构在算法治理、系统性能、数据记录、持续监测规范使用智能技术。美国联邦贸易委员会发布的备忘录要求各企业在人工智能的使用中确保透明、公平、可解释性和稳健性。美国国防部发布的《负责任的人工智能

备忘录》,为第三方开发者开发军用人工智能产品提供了清晰高效的评估和查询流程。美国平等就业机会委员会也曾发起一项致力于消解就业领域算法偏见的倡议,以确保人工智能和其他用于招聘和就业决策的新兴工具使用符合民权法律[11]。

3. 欧盟实践

2018 年 4 月 25 日,欧盟委员会发布了《欧盟人工智能》政策文件,提出了以人为本的人工智能发展理念。作为人工智能的推动者,欧盟也在世界范围内发布了首部人工智能法案。2021 年 4 月 21 日,欧盟发布了《人工智能法》提案,该法案将人工智能应用的风险划分为四种,提出了在人工智能产品上市之前,应对产品进行相应风险评估,不可接受风险不能上市,高风险要有合适的监管,最大限度地缩小风险扩散的范围,要求有限风险的产品应该对公众透明化,对极低风险产品的使用法案不进行干预。作为一项具有权威性的人工智能监管法案,该法案为人工智能治理的可执行性提供了法律参照[11]。

三、人工智能治理的新理念

随着新一代人工智能技术的快速发展,国内外学者对以人工智能为代表的新兴技术治理进行了广泛讨论,逐渐形成了多种治理理念与模式,比较前沿和新型的治理理念主要有以下几种。

1. 敏捷治理

敏捷治理理论最初是在软件开发领域提出的,强调软件开发过程的互动,加强与客户合作,通过工具创新和迭代缩短软件开发周期,以应对快速变化的环境与需求的一种理论。历经多年的发展变化,敏捷治理从软件开发领域发展到多学科应用,其核心理念“以敏捷的态度回应快速变化的环境”得到了诸多学者的认可与发展。2018 年的世界经济论坛发布了标题为《敏捷治理:第四次工业革命时代政策制定的重构》的治理白皮书,将敏捷治理正式引入政府治理领域[12]。该白皮书提出,敏捷治理是一种自适应、以人为本与具有包容性和可持续性的决策过程,其运用的手段应具有灵活性、适应性和柔韧性,敏捷不仅要求更快速的反应与治理行为,同时还要对政策流程进行全新的设计。

在实践方面,敏捷治理首先在西方国家得到了政府的采纳。美国政府采用敏捷的方法与理念,在政府内部通过敏捷治理改进电子政府的效率与服务,从治理流程、服务内容、价值交付等方面进行敏捷化变革。英国数字服务小组(GDS)以敏捷治理理念为指导,建立起敏捷服务社区,通过建立在线服务平台,对社区居民的需求进行迅速的回应,为其提供高效高质的社会服务。第四次工业革命中,以大数据、人工智能、区块链为代表的新兴技术改变了人类社会传统的信息沟通方式,给政府监管、风险控制带来了巨大挑战。在此背景下,敏捷治理被认为是回应新兴技术治理的理想方式。敏捷治理是指应对不断变化的新兴技术采取同等迅速变化的、灵活的、适应性的手段进行治理,敏捷治理通过对快速变化的新兴产业进行治理从而推动其进行创新。

敏捷治理具有如下特征:其一,时间敏感性。敏捷治理需要对快速变化的新技术变化采取同等迅速的手段进行监管和治理,同时进行持续性的准备以应对未知的风险,并在不断变化与技术更迭中实现学习与创新。其二,参与广泛性。敏捷治理强调多方利益主体之间的协同合作,通过让更多的利益主体参与到政策制定与治理过程中来,加强多主体间的协调与合作,将企业的创新活力融入政府治理中,以多方利益平衡机制增进政策的有效性与可持续性。其三,政策工具的更迭与创新。新技术发展为政府治理工作的开展所采取的方法与工具提出了更高的要求,传统的治理工具具有时滞性高、

治理力度大的特征,敏捷治理要求政府不断更新治理工具,以轻柔、灵活的工具为主,以方向修正与抽象指导为主要治理方式,以互动式监管促进产业创新。

2. 韧性治理

“韧性”一词来源于拉丁词语“Resiliens”,但直到 19 世纪初,随着弹性力学基本理论的确立,“韧性”一词才被赋予“压缩后恢复原状能力”的内涵。20 世纪 50 年代,西方心理学引入“韧性”的概念来反映人们遭受精神创伤后的恢复情况。20 世纪 70 年代,生态学家 Holling 首次使用“韧性”一词描述干扰过后生态、资源系统的自我修复能力,并认为这种性质不仅意味着系统能恢复到原初平衡状态,更重要的是可以向新的平衡状态转化[13]。

近年来,国内外专家学者纷纷将韧性治理引入到公共管理领域,韧性治理是指在系统分析脆弱性风险的基础上,运用更具包容性、冗余度、感知力与应变力的治理手段,通过事前预警规避、事中响应调整、事后迅速恢复和优化的系统流程来提升风险治理水平[14]。韧性治理的理念逐渐融入国家治理、城市建设、社区治理、产业治理的各个方面和各项具体行动之中。在治理过程中,以人工智能、大数据、云计算为代表的新兴技术,一方面,可以作为治理工具赋能韧性治理,协助进行智能社会治理、智慧城市治理,防范和化解城市和生活中存在的各种风险和矛盾,提升治理的效率和效能;另一方面,人工智能等新兴技术也作为治理对象存在,治理者对人工智能存在的风险进行识别和预测并制定治理预案,对已经出现的风险和问题进行有效干预和控制。只有依法和适度地应用人工智能,才能让人工智能在韧性治理过程中实现治理工具和治理对象的统一。

3. 负责任创新和治理

2003 年,德国学者海斯托姆首次提出了“负责任创新”的概念,随着新兴技术的不断发展,各国政府和学者也在不断思考“创新创造”和“治理”之间的辩证关系,“负责任创新”的内涵被不断拓展。2011 年,欧盟政策委员会发布了《地平线 2020 框架计划》,其中明确提出“负责任创新”这一概念,并将其作为欧盟发展战略的重要内容。该计划指出,负责任创新的基本要素包括社会利益、道德伦理可接受程度及风险管理等,强调将科技进步合理地嵌入社会发展进程,引导科技创新的过程与产品实现伦理可接受、发展可持续和社会满意,实现科技创新面向未来发展的集体管理[15]。

负责任创新聚焦新兴技术治理,推动了传统的以风险议题为核心的技术治理模式走向创新行为的责任塑造模式,形成了目标设定、行动主体参与、价值准则协调、过程响应、制度建构五个维度[15],逐渐成为新兴技术治理的主流治理模式和理念之一。欧盟针对新兴技术治理的责任式创新范式的提出,得到全球多个国家如美国、英国、印度、韩国等的积极响应。中国随即也展开了对负责任创新的新兴技术治理的研究与实践,2019 年 6 月 17 日,我国发布《新一代人工智能治理原则》,其中明确提出了“发展负责任的人工智能”[15]。

四、启示与展望

人工智能作为全球前沿的创新领域,其产生的颠覆性突破往往会带动其他产业的发展甚至引发整个社会的综合性变革。因此,对于人工智能的发展和治理,不仅要充分释放“制”的优势,更要做好“治”的文章,充分发挥各治理主体的作用,需要政府发挥统领监督引导作用,整合企业、科研机构、高校的各种资源,实现政府、企业、高校、科研机构的协同发力。

人工智能治理机构应从技术上致力于提高人工智能技术的可解释性、信任度、评估监测

等,同时也应开展人工智能行为科学和伦理等问题研究,研制多层次的人工智能治理框架,研究人工智能立法规划和制定人工智能突发事件应急方案。针对人工智能存在的“黑箱”问题和不可解释性,应努力提高算法透明度,如规定在某些情形下公开人工智能系统或应用代码,规定公众或受人工智能自动决策影响的用户有权知道人工智能算法背后的基本逻辑或运算标准,用户在受到不公正对待时有权获得救济并对相关责任主体追究责任。

参考文献:

[1] RAPAPORT W J. What is artificial intelligence? [J]. Journal of Artificial General Intelligence, 2022(2):52.

[2] MAAS M. Artificial intelligence governance under change foundations, facets, frameworks [D]. Denmark:University of Copenhagen,2020.

[3] 庞祯敬,薛澜. 人工智能治理:认知逻辑与范式超越[J]. 科学学与科学技术管理,2022(9):3.

[4] 闫德利. 2016 年人工智能产业发展综述[J]. 互联网天地,2017(2):22.

[5] 石琳娜,陈劲. 数字经济推动实现共同富裕的机理与路径研究[EB/OL]. (2022 - 11 - 30)[2022 - 12 - 01]. https://kns.cnki.net/kcms/detail//42.1224.G3.20221130.1432.003.html.

[6] 鲁传颖,张璐瑶. 人工智能的安全风险及治理模式探索[J]. 国家安全研究,2022(4):84.

[7] 洪永淼,汪寿阳. 人工智能新近发展及其对经济学研究范式的影响[J]. 中国科学院院刊,2023(3):353.

[8] LIU T. Algorithm-dependent generalization bounds for multi-task learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016(39):227.

[9] 梁正,张辉. 构建平衡包容的人工智能治理体系[J]. 中国发展观察,2022(12):44.

[10] 刘露,杨晓雷,高文. 我国人工智能伦理监管需求分析及对策研究[J]. 中国工程科学,2021(3):106.

[11] 段家欣,赵瑜. 全球人工智能治理的核心议题与规制趋势[J]. 声屏世界,2022(12):5.

[12] 曹海军,侯甜甜. 敏捷赋能视角下的数字政府建设:实践缘起与理论建构[J]. 吉林大学社会科学学报,2021(6):170.

[13] HOLLING C S. Resilience and stability of ecological systems[J]. Annual Review of Ecology and Systematics,1973(4):1.

[14] 翟绍果,张星. 从脆弱性治理到韧性治理:中国贫困治理的议题转换、范式转变与政策转型[J]. 山东社会科学,2021(1):76.

[15] 梅亮,臧树伟,张娜娜. 新兴技术治理:责任式创新视角的系统性评述[J]. 科学学研究,2021(12):2113.

[责任编辑:王天笑]

引用格式:石琳娜. 人工智能发展的风险与治理[J]. 郑州轻工业大学学报(社会科学版),2023,24(2):56 - 62.