人工智能风险的试探性治理:概念框架与案例解析

李　冲[1],李　霞[2]

(1. 大连理工大学高等教育研究院,辽宁　大连　116024;2. 山西师范大学教育科学学院,山西　太原　030031)

摘　要:有效治理人工智能风险与发挥其溢出带动的“头雁效应”同等重要。工具主义和结构主义两种竞争视角下的治理路径,都难以作为人工智能水平领域治理的主导逻辑。采用“技术双向性”的研究思路,构建“技术—制度”协同演化的试探性治理概念框架,以深度伪造技术风险治理为典型案例进行探索性解析。研究发现,人工智能风险治理是一个技制迭代协同演化的过程。具体而言,技术治理由“被动检测”逐渐发展到“主动检测”;制度治理从“间接监管”逐渐转向“直接监管”;技制迭代发展由“治标型协调”演化形成“治本型协同”。有效治理人工智能风险需要树立弹性思维观念,把历时分析和辨证思考引入治理过程,及时、合理地配置技术工具和制度工具。

关键词:人工智能;风险治理;试探性治理;案例研究

中图分类号:G310　　**文献标识码**:A　　**文章编号**:1005-0566(2024)04-0091-11

Tentative governance of artificial intelligence risk: Conceptual framework and case analysis

LI Chong[1], LI Xia[2]

(1. *Institute of Higher Education*, *Dalian University of Technology*, *Dalian* 116024, *China*;
2. *College of Educational Science*, *Shanxi Normal University*, *Taiyuan* 030031, *China*)

Abstract: The effective management of artificial intelligence risk is as important as the “head goose effect” driven by its spillover. The governance paths under the competitive perspective of instrumentalism and structuralism is difficult to be the leading logic of governance in the field of artificial intelligence. This paper adopts the research idea of “technology bi-directionality”, constructs a tentative conceptual framework of “technology—institution” co-evolution, and takes deep forgery technology risk governance as a typical case analysis. It is found that artificial intelligence risk governance is a process of iterative development and co-evolution of technological governance and institutional governance. Specifically, technological governance has gradually developed from “passive detection” to “active detection”; institutional governance has gradually changed from “indirect supervision” to “direct supervision”; iterative development of technical system has evolved from “temporary coordination” to “permanent coordination”. Effective governance of artificial intelligence risks requires establishing a flexible thinking logic, introducing diachronic analysis and dialectical thinking into the governance process, and timely and reasonable allocation of technical and institutional tools.

Key words: artificial intelligence; risk governance; tentative governance; case study

收稿日期:2023-12-12　　**修回日期**:2024-03-06

基金项目:国家自然科学基金面上项目(71974027);山西省哲学社会科学规划课题项目(2022YJ061)。

作者简介:李冲(1974—),男,吉林农安人,大连理工大学高等教育研究院教授、博士生导师,研究方向为科技政策、教育政策。通信作者:李霞。

以 Chat GPT 为代表的新一代人工智能技术快速迭代发展,引发社会生产、生活方式和组织形式重大变革。作为“推动科技跨越发展、产业优化升级、生产力整体跃升的重要战略资源”[1],人工智能可为人类社会带来巨大福祉,同时也会引发诸多风险,危及到个人隐私、劳动就业、国家安全和公共安全,造成广泛的信任危机、伦理危机和社会危机。

当前,对人工智能风险治理的研究主要集中在垂直领域和水平领域两个方向。综合来看,现有研究成果丰硕,尤其是在强调定制化需求和具体解决方案的垂直细分领域,如医疗保健、金融、零售、交通运输等行业,相关研究十分具体深入。但是在横跨多个行业,强调通用性和实用性的水平领域,例如数据隐私保护、算法规制、场景应用等,相关研究进展缓慢,而此类问题正是社会普遍关注的重点和难点。目前,水平领域的研究主要有工具主义和结构主义两种竞争性的视角。从工具主义视角出发的治理思路聚焦人工智能的技术赋能特征,强调人工智能的工具属性,因此“以技治技”的治理逻辑较为明显。以中国知网数据库的“人工智能技术治理”主题词检索可以发现,聚类主要集中在“技治主义”“实时技术评估”“敏捷治理”等。从结构主义视角出发的治理思路侧重社会对技术变革的嵌入性,强调制度引导、形塑技术的发展方向,因此“以制适技”的治理逻辑比较突出。相关研究认为,发展人工智能技术不仅要注重审慎原则(precautionary principle),而且还要走出受控的科学实验室,介入非原型的不受控的“社会实验室”“生活实验室”等[2]。

从当前人工智能技术发展的态势以及风险治理需求的角度看,上述两种治理逻辑都具有广阔的适用场景和广泛的研究空间,但是都难以作为人工智能水平领域治理的主导逻辑。一方面,人工智能具有广泛的通用性、覆盖性和溢出带动性,因此风险治理自然也不宜从单一维度简单归结为“以技治技”抑或“以制适技”,而应从整体主义视角出发,将技术和制度的关系纳入整个治理场域进行辨证思考。另一方面,社会风险的生成也是一个主观、客观要素交织作用的历时性问题,因此理应把人工智能风险治理作为一个动态变化的系统工程,从技术和制度互动的角度探讨其演化过程、辨证关系与关键机制。基于上述分析,本文把人工智能技术发展及其社会风险生成作为先验的治理前提,采用“技术双向性”的研究思路,将工具维的技术赋能与制度维的适应调试相互链接,尝试构建一个“技术—制度”协同演化的试探性治理框架,并基于此探讨当“科林格里奇困境”出现端倪时,工具主义治理的技术逻辑和结构主义治理的制度逻辑之间的互动机制、平衡关系与演化机理。

一、理论回顾与研究框架

(一)工具主义视角下的人工智能风险治理路径

工具主义视角下的人工智能治理强调风险的内生性,主要的关注点是人工智能快速发展引发数据、算法和应用场景的治理议题,强调从科学意义上解决开发、应用中的技术性能、创新管理和技术影响等新的科学问题。

第一,基于技术性能优化的质量管理路径。由于人工智能偏离预期结果的科学原因主要取决于数据与算法,因此这一治理思路试图通过制定“标准化方案”实现数据无偏失、算法无缺陷,达到人工智能技术“无害化”目的[3]。一是,强调制定数据质量标准和流程,如制定数据质量评估指标,建立数据质量监控机制等,以确保数据的准确性、完整性、一致性和真实性。二是,规范算法开发流程,建立算法“黑盒”测试环境、技术产品和测试标准,实现防范人工智能技术风险的“关口前移”。如 2018 年欧盟出台的《通用数据保护条例》规定,运用自动化算法的企业要能够保证算法具有透明性和可解释性。2023 年我国七部门公布的《生成式人工智能服务管理暂行办法》要求,人工智能使用的数据和基础模型要具有合法来源,训练数据要增强真实性、准确性、客观性、多样性。

第二,基于技术创新主体责任的过程管理路径。由于数据获取、算法开发与主体选择密切相关,技术创新主体基于不同的价值、利益和专业视

角处理、开发和应用人工智能技术,认知逻辑的“框架前提”差异直接决定了人工智能技术发展偏离预期结果的程度。因此,这一治理思路在强调提高数据、算法质量以及开发检测技术的同时,进一步强化创新主体责任,加强对人工智能技术研发与市场化的过程管理。如2022年中国三部门发布的《互联网信息服务深度合成管理规定》要求,深度合成服务提供者和技术支持者要加强训练数据和算法管理,保障训练数据和算法安全。对使用算法服务生成或者编辑的信息内容,要采取技术措施添加不影响用户使用的标识,对可能导致公众混淆或者误认的,要在生成或者编辑的信息内容的合理位置、区域进行显著标识,以向公众提示深度合成情况。《生成式人工智能服务管理暂行办法》也要求,生成式人工智能技术研发过程中进行数据标注,提供者要制定清晰、具体、可操作的标注规则;要开展数据标注质量评估,抽样核验标注内容的准确性。

第三,基于技术特征识别的风险分类管理路径。由于人工智能技术支撑的通用性和影响的跨域性特征,这一治理思路主要以人工智能赋能场景为依据,对人工智能应用在不同行业(如医疗、交通、金融等)风险的具体表现进行类型学划分并实施分类治理。这一治理思路的本质是沿袭传统的风险治理模式,将人工智能的不确定性风险治理,化约为“发生损害概率和损害严重程度组合”的确定性风险治理。如2016年欧盟《通用数据保护条例》(简称GDPR)和2023年《人工智能法案》,将人工智能风险分为不可接受风险、高风险、限制性风险和最小风险四种类型,强调根据人工智能系统可能产生风险的强度和范围来确定治理规则的类型和内容。与之不同的是,我国全国信息安全标准化技术委员会2021年发布的《网络安全标注实践指南——人工智能伦理安全风险防范指引》没有采用这一划分方式,而是按照人工智能技术相关风险的性质分为“失控性风险”“社会性风险”“侵权性风险”“歧视性风险”“责任性风险”五类,强调在研究开发、设计制造、部署应用和用户使用等环节识别、防范、管控人工智能安全风险。

(二)结构主义视角下的人工智能治理路径

结构主义视角下的人工智能治理强调风险的外部性,主要的关注点是人工智能快速发展引发技术利维坦、数字鸿沟、信息茧房、马太效应和伦理困境,如此等等问题带来了新的社会矛盾,强调从伦理规范、监管规制和后工业社会角度解决不确定性、复杂性和模糊性等新的治理议题。

第一,基于伦理建构的规范性路径。由于技术风险外部性的成因与风险的社会建构性密切相关,这一治理思路把人工智能风险的根源归因于经济结构或社会秩序变化所带来的权力、权利和利益分配格局的失序。因此这一治理思路从人类良善价值的角度出发,强调通过利益相关者的互动和理解达成对伦理、道德、意义等价值的“最大公约数”,引导人工智能技术向善发展。如2019年经济与合作组织发布的《确定OECD的人工智能发展原则》报告,提出了包容、可持续发展与福祉,以人为本的价值观和公平,透明和可解释性,健壮性和安全性,可问责制五项原则;同年,欧盟发布的《可信赖人工智能道德准则》报告,提出了人类代理和监督,技术强大性和安全性,隐私和数据治理,透明度,多样性,非歧视性和公平性,社会和环境福祉以及问责制七项原则。2023年10月,我国中央网信办发布的《全球人工智能治理倡议》,围绕人工智能发展、安全、治理三方面,系统阐述了坚持伦理先行的人工智能治理中国方案。

第二,基于法律监管的规制性路径。通过制定新的或修订已有法律、法规用以监管人工智能理论发展、技术创新、数据归集、算法迭代等,是目前世界各国应对新兴技术风险的常规做法。由于人工智能风险的不确定性、利益的交织性和社会影响的复杂性等特征,这一治理思路主要形成了一种治理型监管(governance-based regulation)的模式,亦以监管权的开放协同、监管方式的多元融合、监管措施的兼容配适为核心特征的新型监管范式[4]。如中国、欧盟和美国面对人工智能发展可能引发的各种社会风险,基本上都没有采用侧重于防范风险的扩张性“强监管”思路,而是采用了鼓励创新发展的谦抑性“弱监管”思路。相对而

言,欧盟表现出的“严”监管特征源于其人工智能技术创新落后于中国和美国,旨在通过制定技术标准和行为规则实现其赢得技术话语权的目的;而美国则由于技术创新的领先地位,表现出企业友好型的“宽”监管特征,目的是在于维护其科技领先地位。中国则表现出“严”“宽”相济的平衡性特征,这主要是源于中国发展人工智能技术的根本目的在于促进数字经济高质量发展,提高自主技术创新能力和水平,由此带动产业升级转型和增进民生福祉。

第三,基于后工业社会的反思性路径。由于包括人工智能在内的新兴技术突破了传统技术风险影响的边界,西方国家主导的现代性存在“有组织地不负责任”[5]、“人为制造的风险”[6]以及“风险的内部化”等天然缺陷,因此这一治理思路强调要根除“以资本为中心”的强制逻辑和同化倾向,通过重塑技术与资本的“良善”关系,防止人工智能偏离“以人为本”的异化。如贝克认为资本逻辑主导的工业社会潜藏着意义深远的制度性危机:“财富在上层聚集,风险在下层聚集”,而逆转这一危机的唯一路径就是以风险的生产逻辑取代财富生产的逻辑[7]。在治理实践中,这一思路从最初的审慎原则发展成为了负责任的研究与创新(responsible research and innovation,RRI)。2021 年,欧盟启动的研发与创新框架计划——欧洲地平线(Horizon Europe),明确将 RRI 理念贯彻其中,同时还希望将这一理念在全球范围内推广,从而建立起世界性的 RRI 体系。2019 年,中国发布的《新一代人工智能治理原则——发展负责任的人工智能》报告率先采用了这一治理理念。2023 年 10 月 18 日,中国发布的《全球人工智能治理倡议》拓展了这一理论框架,进一步强调共同做好风险防范,形成具有广泛共识的人工智能治理框架和标准规范,不断提升人工智能技术的安全性、可靠性、可控性、公平性。

(三)试探性治理:一个技术与制度协同演化的框架

围绕人工智能的内生性和外部性问题,工具主义视角的治理路径开发出了预见性治理、实验主义治理、敏捷性治理等框架;结构主义视角的治理路径开发出了伦理治理、自反性治理、适应性治理等框架。总体而言,工具主义和结构主义视角的不同思路分别在人工智能治理的垂直领域和水平领域具有较好的适恰性,但是都难以成为一个契合性更强的总体治理框架。根本性的问题在于:一方面,技术塑造着我们,深刻改变了我们的生产生活,拓宽着我们的思维方式和认知水平;另一方面,技术发展是社会选择的结果,我们的社会(包括习惯、价值、组织、思想和风俗等)决定了技术发展的轨迹与状况,也以独特的方式塑造了技术。因此,本文采用“技术双向性”视角,从技术与制度协同演化的关系处着眼,建立一个辩证的人工智能试探性治理框架。

所谓试探性治理(tentative governance),指的是与确定性治理(definitive governance)过程分解式(风险识别、评估、管理和反馈)的线性治理方式相对应,尝试通过循环迭代的方式来管理复杂性、不确定性问题的一种治理模式。正如斯蒂芬·库尔曼(Stefan Kuhlmann)等人所定义的那样,当治理被设计、实践、行使或发展为一个动态过程,并以非最终化的方式管理相互依赖和突发事件时,这种治理模式就是“试探性”的。这里的“试探性”表现为谨慎、初步、灵活和渐进地处置问题,而非果敢、坚定地寻求最终化的方案[8]。试探性治理并不是一个排斥其他治理模式的独特类型,它与适应性治理、预期治理等模式既具有一定的区别,又存在广泛、深入的联系(见表 1)。

表 1　试探性治理与其他治理模式的比较

治理模式	治理目标	治理工具	决策方式	治理范式
自反性治理	重塑问题解决程序	认知和制度的现代性反思	社会话语、政治过程	探索性场景构建
预见性治理	建立行动者的远见、参与和整合能力	实时技术评估	公众参与	增强治理过程的互动性和有效性
适应性治理	管理快速变化和高不确定性的环境	社会实验、社会学习	公民参与、市场干预	社会生态系统重组和调整
实验性治理	提升治理实践的科学性和有效性	政策实验	参与式合作	实验推广
敏捷性治理	提高决策适应能力	政策实验室、监管沙盒	利益相关者参与	持续监管机制创建
试探性治理	创造探索和学习空间	实验、经验评估	开放式参与	设计和践行审慎、灵活、渐进的动态治理机制与路径

与自反性治理相比,试探性治理比自反性治理涵义更为广泛,更加强调是否以及如何实现自反性治理;与预见性治理相比,试探性治理更强调开发未来可能的治理方案;与适应性治理相比,试探性治理更强调通过启发性过程寻找更具适应性和柔性的治理方案;与实验性治理相比,试探性治理更强调不确定性而不是目标设定;与敏捷性治理相比,试探性治理并不单纯强调持续的监管,而且还预留了技术主动调试的空间[9]。从人工智能的技术赋能和社会嵌入等特征看,人工智能风险治理不仅需要审慎地"关口前移",而且还要在技术市场化和风险社会化的动态过程中,连续、灵活、渐进地不断调整,最终才能实现技术与制度的适应和匹配。因此,与其他治理模式相比,使用试探性治理模式的思路来构建人工智能风险治理的总体框架更为适恰。如图1所示,人工智能风险治理的试探性框架主要由技术赋能、社会嵌入及其互动迭代3个部分构成。

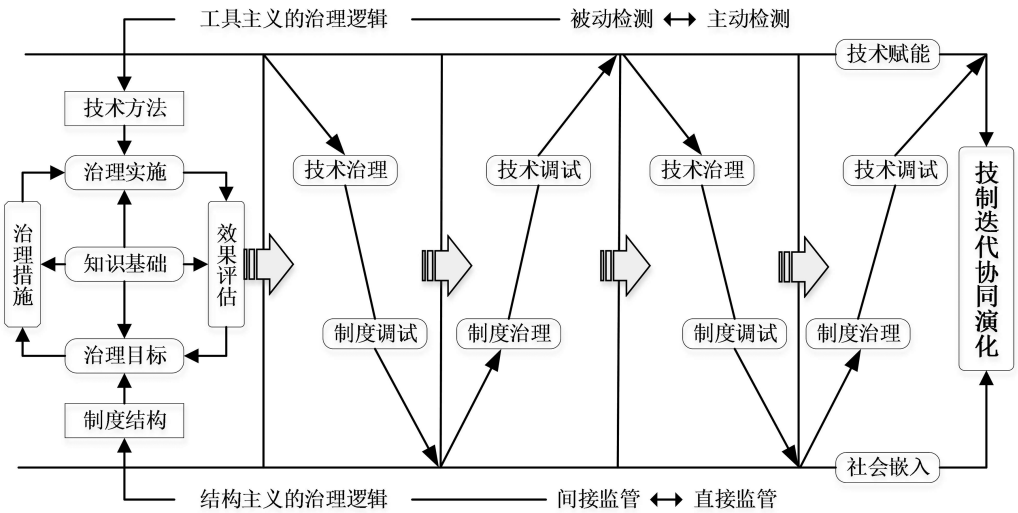

图1　人工智能技术风险试探性治理的协同演化框架

二、研究方法、案例选择与资料收集

(一)研究方法

由于人工智能是一项具有"乌卡特征"①的技术,在风险治理问题上远未达成共识性观点,因此本文选择案例研究方法。具体将理论分析与实践经验相结合,通过搜集典型案例资料,结合研究主题和案例本身对事件过程进行分析、讨论、提炼与总结,进而回答"为什么(why)"和"如何做(how)"的问题。目前,案例研究法主要划分为单案例研究和多案例研究两种。单案例研究适用于解析极端案例,对处于特殊情境、缺乏理论阐释基础的单独事件给予情境化、个性化的分析;多案例研究则是对兼具多样化、普遍性与代表性特征的案例进行解构讨论,通过选取多个典型案例,以一定的理论基础解析其内在规律,探讨普遍适应性[10]。由于人工智能风险具有普遍性和延伸性的特点,因此本文采用多案例研究方法。

根据案例选择的理论抽样和条件类化逻辑,多案例研究的理想案例数量是3~7个[11]。考虑到案例数量的基本要求与增加案例个数导致的研究成本增加,本文选取3个案例作为分析对象。在具体案例方面,考虑到世界各国的人工智能技术发展水平差异较大,而美国、欧盟和中国是人工智能第三次浪潮的主要引领者,因此选取美国、欧盟和中国的人工智能治理案例作为分析对象[12]。

(二)案例选择

人工智能是包括计算机视觉、语音识别、自然语言处理、机器学习、大数据等多种技术的统称。

① 乌卡(VUKA)是volatile(波动性)、uncertainty(不确定性)、complexity(复杂性)和ambiguity(模糊性)的集合,主要用以描述无法理论推演、无法逻辑分析、没有数据证明、甚至无法经验总结,而是一切处于实时未知中的混沌状态。

考虑到案例的典型性,本文选择深度伪造(deepfake)技术作为代表性案例。深度伪造技术是利用“生成式对抗网络”(generative adversarial networks,GAN)机器学习的深度学习算法,实现语音模拟、人脸合成、视频生成等人工智能的一个技术分支[13]。这一技术生动诠释了“眼见不一定为实”的新现象、新问题。最初,深度伪造技术更多地用于娱乐活动,如整蛊朋友或调侃明星等,后来深度伪造逐渐“走向政坛”,为国家间的政治抹黑、经济犯罪甚至恐怖主义行动等提供了新工具。比如,2023 年的一段“印度总理莫迪跳传统舞蹈”视频在网上疯传,莫迪本人回应表示此视频是由人工智能合成的深度伪造视频。2019 年,德国发生了犯罪者利用深度伪造技术模仿某能源公司首席执行官声音实施诈骗的事件。2018 年,美国发生了伪造前总统奥巴马调侃时任总统特朗普的视频,在推特上获得了 200 多万次播放和 5 万多个点赞。截至目前为止,深度伪造技术在我国尚未发生严重的滥用事件。不过,研究者普遍认为,深度伪造技术一旦被滥用,将会引发社会忧虑和信任危机,将给个人或企业带来损害,严重的还会威胁国家安全和公共安全。总之,深度伪造技术的出现使人类进入伪真相时代,滥用该技术将消解人类视觉的客观性和真实世界的符号规制,不仅会侵犯公民人格权、财产权、知识产权等合法权益,还会引发心理迷思、社会失范、信任危机等,严重的甚至可能导致社会共识崩塌[14]。

(三)资料收集

考虑到数据三角验证要求,本文的数据来源主要包括研究论文、专题网站报告和新闻报道。①研究论文。主题研究论文是对特定议题的深入分析与讨论,不仅能提供理论借鉴,也间接地提供了一些档案记录数据,具有很强的专业价值。本文主要在中国知网和 WoS 等数据库通过检索深度伪造的主题论文进行案例数据收集。②专题网站报告。专题网站报告主要包括官方网站和专业组织网站发布的相关报告。互联网的发展普及提升了深度伪造相关信息数据的公开性与透明度,官方网站和专业组织网站发布的相关报告更加集中地汇集了深度伪造议题的统计数据,使得资料收集更加便利和全面。③新闻报道。新闻报道是对新事物的及时准确传播,具有强实时性特点,因此,来自诸如搜狐网、央视网、华尔街日报等的相关新闻报道数据对深度伪造技术研究也有较强的适用性。

三、案例分析:人工智能深度伪造技术风险的试探性治理

(一)技术赋能与技术治理:从被动检测到主动检测

在技术发展史上,一种新技术在发展的早期阶段往往存在乐观主义和悲观主义两个对立的认识阵营[15]。与之不同的是,新一代人工智能风险在早期阶段就形成了比较一致的认识。比尔·盖茨、斯蒂芬·霍金、沃伦·巴菲特、埃隆·马斯克等都曾发出警告,必须高度重视人工智能可能带来的不确定性影响。深度伪造作为一种具有丰富应用场景的人工智能技术,可以为不同行业和领域提供增值服务,但同时也会冲击法律和社会伦理、侵犯个人隐私、危害国家安全和社会安全。世界各国政府在着力推动深度伪造技术赋能作用的同时,在风险的技术治理方面,除对数据、算法等提出一般性的质量和管理要求之外,一个比较显著的特征是高度关注深度伪造技术的发展变化,进而开发相应的伪造检测技术[16]。

在深度伪造技术发展的早期阶段,美国、欧盟和中国主要是采用“你攻我防”的方式开发“被动检测”技术。具体使用的技术如卷积神经网络(convolutional neural network,CNN)和复合模型缩放(efficientnets)等伪造特征识别与提取的鉴别技术[17]。2018 年,美国国防部高级研究计划局(DARPA)推出了全球首款变脸检测工具。2019 年,北京瑞莱智慧科技有限公司(RealAI)与清华大学人工智能研究院推出了基于多种攻防算法和模型的 RealSafe 对抗攻防平台[18]。随着深度伪造技术的不断发展,检测技术开发也由伪造特征识别与提取的被动检测,逐渐转向基于深度学习的特征鉴别和基于融合边界的伪造检测等“主动检测”技术。如 FaceX-Ray 模型、单中心损失检测(single center

loss,SCL)、混合图像检测(self-blended images,SBI)等[19]。2019 年,美国加州大学伯克利分校和南加州大学合作研发了识别生物标签的人工智能系统,并构建高度个人化的“软生物识别指标”体系。这个系统能够通过面部表情和头部动作等生物标签解析视频是否被“伪造”[20]。同年,斯坦福大学推出基于区块链技术追踪图像与视频真实性源头的伪造鉴别技术[21]。2020 年,微软公司推出视频认证器(video authenticator)检测技术,并通过了人脸视频深度伪造检测挑战赛(DeepFake Detection Challenge,DFDC)数据集测试。同年,北京大学与微软亚洲研究院联合推出了面部假图像检测工具“Face X-Ray”,检测精准度可达 95% 以上。2021 年,瑞莱智慧科技有限公司推出了基于贝叶斯深度学习、多特征融合和多任务学习等方法的 DeepReal 深度合成内容检测平台[22]。

(二)社会嵌入与制度治理:从间接监管到直接监管

人工智能是一种颠覆性技术,不仅能够改变物质生产方式,而且也会影响社会生活的方方面面。制度通过刚性约束和弹性激励,把社会因素嵌入技术发展过程中,以降低社会风险发生。在制度治理方面,除了一般性的行为规范和伦理要求之外,美国、欧盟和中国对深度伪造技术均表现出由间接监管转向直接监管的特征,同时由于各国社会治理体制的差异,制度治理的主导方式和路径略有不同。

1. 美国:以权益保护为主导的上下分治型制度治理

美国是人工智能浪潮的发源地,为维持其技术、产业优势和领导地位,美国把人工智能风险治理更多地赋权给谷歌、微软等科技巨头以及地方政府,只有在涉及到重大利益关系时,联邦政府才将介入。2019 年之前,美国对深度伪造的制度治理主要采用自下而上的间接监管,联邦政府较少出台规制性法案。如早在 2009 年得克萨斯州就出台了《生物识别信息获取使用法》,直到 2018 年 12 月,联邦参议院才提出《2018 年恶意伪造禁令法案》。由于深度伪造技术已被应用到政治、经济和社会等各领域,并产生一系列消极影响,2019 年之后,美国对深度伪造的制度治理开始由自下而上的间接监管转向联邦政府自上而下的直接监管。相关数据表明,在深度伪造社会风险发生的短时期内,联邦立法机构提出了多项法案[23]。2019 年 6 月 12 日,众议院推出《深度伪造责任法案》,众议院情报委员会主席亚当·希夫(Adam Schiff)于 13 日围绕深度伪造技术举行了首次听证会,督促联邦和各州立法机构积极落实应对行动。2019 年 6 月 28 日,参议院和众议院共同推出《2019 年“深度伪造”报告法案》。这些法案明确规定了技术使用者的义务、侵权者的刑事责任或行政责任及惩罚措施等,为深度伪造技术的直接监管奠定了基础[24]。在一系列深度伪造事件和学界呼吁影响下,美国各州分别制定规制更加严厉的相关法案。如加利福尼亚州分别于 2019 年 2 月和 10 月颁布《犯罪:欺骗性记录》和《选举:欺骗性的音频和视觉媒体》两项法案,严禁使用深度伪造技术传播色情和政治助选。弗吉尼亚州 2019 年 7 月出台《非法传播或出售他人影像》法案,对深度伪造侵权行为明确规定了法律责任,具体可判处最高一年的监禁和 2 500 美元的罚款[25]。

2. 欧盟:以行动框架为主导的集中型制度治理

欧盟一直将人工智能视为缩小新一代技术差距与拉动经济增长的关键领域。不过,由于欧盟没有强有力的中央政府,同时也缺少如同谷歌、微软等大型科技企业,因此欧盟的人工智能制度治理强调在欧盟整体与成员国层面不断加强立法与合作,以通过行动框架构建巩固其在技术规则上的主导地位。针对人工智能技术,欧盟委员会出台了一系列相关法律、法规和政策文件,如《通用数据保护条例》《可信赖人工智能伦理指引》等。在深度伪造技术发展的早期阶段,欧盟主要采用的是综合性的间接监管,法案、政策等规定的内容相对较为宽泛,主要是将深度伪造技术纳入生物信息保护和不实信息处置框架中进行综合治理。2018 年 5 月,欧盟发布的《通用数据保护条例》第 9 条规定,除非满足特殊条件,对于识别自然人的

生物性数据应当禁止处理[26]。2018 年 6 月,欧盟理事会通过的《欧盟不实信息实践准则》,要求行业进行自我规制、合作研究、建立评估审核机制、提升公民媒体识读素养等[27]。随着深度伪造技术风险不断升高,欧盟开始采取自上而下的方式进行直接监管,明确将深度伪造技术纳入人工智能规制框架中进行全周期治理。2019 年 4 月,欧盟执行委员会 AI 高级专家工作组(High-Level Expert Group on AI)提出了《可信赖人工智能伦理指引》,从全生命周期角度对深度伪造技术进行了规定,并指出合法性、伦理性和鲁棒性(robust)是防范其风险的关键。2021 年 5 月,欧盟委员会公布《加强<反虚假信息行为守则>的指南》,提出互联公司采取有效应对措施,监管机构扩大核查覆盖面,增加研究人员获取数据的机会等规定。在此基础上,2022 年 6 月出台了更加严格的强化版指南,要求并获得了 34 个相关平台、科技企业与公民组织的遵守承诺[28]。

3. 中国:以场景应用为主导的部委协作型制度治理

中国发展人工智能技术具有得天独厚的优势。科教兴国战略、人才强国战略、创新驱动发展战略的实施,为人工智能技术发展打下了坚实的社会治理基础。同时,中国的产业结构和社会规模,为人工智能技术发展提供了丰富的应用场景。基于此,中国的人工智能制度治理主要是聚焦于场景应用,以相关部委协作为主导方式。对于深度伪造技术,中国同样也经历一个由间接监管向直接监管转变的过程。间接监管主要涉及强化互联网平台义务和保护个人信息安全等;直接监管包括细化国家与地方的监督管理方式和出台风险防范政策等,但在整个治理过程中这两种规制方式的使用阶段划分并不十分明显。在间接监管方面,2019 年 1 月,中央网信办、工业和信息部、公安部、国家市场监管总局等 4 个部门联合发布了《关于开展 App 违法违规收集使用个人信息专项治理的公告》。2019 年 11 月,网信办、国家广播电视总局等相关主管部门颁布的《网络音视频信息服务管理规定》要求,利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播非真实音视频信息的,应当以显著方式予以标识。在直接监管方面,2019 年 12 月,网信办发布的《网络信息内容生态治理规定》明确要求,不得利用深度学习、虚拟现实等新技术新应用从事法律、行政法规禁止的活动。2020 年 5 月,《民法典人格权编》第 1019 条规定,任何组织或者个人不得以丑化、污损,或者利用信息技术手段伪造等方式侵害他人的肖像权。2022 年 11 月,网信办发布的《互联网信息服务深度合成管理规定》明确要求,任何组织和个人不得利用深度合成服务从事危害国家安全和利益、扰乱经济和社会秩序、侵犯他人合法权益等法律、行政法规禁止的活动,并强调国务院电信主管部门、公安部门、地方网信部门、地方电信主管部门和地方公安部门依据各自职责进行监督、统筹等管理工作。

(三)技制迭代:技术治理与制度治理的协同演化

从美国、欧盟和中国的深度伪造技术治理发展过程上看,3 个国家(地区)的人工智能治理都经历一个技术治理与制度治理的迭代演化过程(见图 2)。整个治理模式特征由“治标型协调”(被动检测—间接监管)发展到“治本型协同”(直接监管—主动检测)。

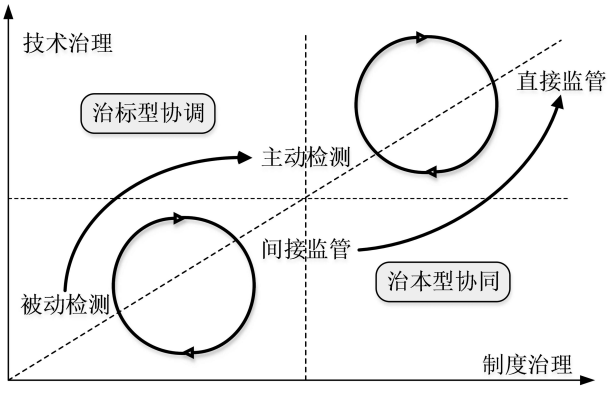

图 2　人工智能深度伪造技术治理与制度治理协同演化

在深度伪造技术发展早期,由于风险和危害尚不明确,美国、欧盟和中国都采用了技术上“被动检测”和制度上“间接监管”的治理方式,技术治理和制度治理的协同度不高。美国政府对深度伪造技术保持积极的态度,强调发挥市场主体在风

险治理中的作用。欧盟将深度伪造纳入虚假信息范畴,协调各成员国参与,要求互联网平台制定行业自律准则、开发检测工具。中国的治理方式也是鼓励创新和防范风险兼顾,伪造检测技术发展基本上与国际水平齐平,监管政策以强化互联网平台义务和保护个人信息安全为主。

随着人工智能技术迭代发展,深度伪造风险事件频发,并且向社会领域逐渐扩散,制度治理则由"间接监管"转变为"直接监管",技术治理也从"被动检测"向"主动检测"转向。技术治理与制度治理的协同演化程度加快、加深。美国在政学两界的共同呼吁下,2019 年联邦政府和各州政府出台了一系列有关深度伪造技术的直接监管法案,检测技术也由伪造特征识别与提取转向"软生物识别指标"等主动检测技术。欧盟同时期也颁布了《可信赖人工智能伦理指引》等一系列直接监管的框架性文件,加强对深度伪造技术的全面监管。中国则是出台了《互联网信息服务深度合成管理规定》等一系列相关政策对深度合成的信息内容、使用方式等进行直接监管,与前期已经施行的《网络安全标准实践指南——人工智能伦理安全风险防范指引》《新一代人工智能伦理规范》等规范性文件,形成了一个完整的制度治理框架。同时,重点研究型大学开始参与深度伪造检测技术的改进与开发,检测技术水平和检测准确率不断提高。

综合来看,美国、欧盟和中国在人工智能治理过程中,技术治理和制度治理都起到了重要作用,并且在互动过程中不断调适与匹配,逐渐由"治标型协调"转变为"治本型协同",形成了技术治理和制度治理迭代发展、协同演化的总体格局。

四、结论与进一步的讨论

(一)基本结论

治理人工智能风险与发挥其"头雁效应"具有同等的重要性。工具主义和结构主义视角下的多种治理模式都难以作为人工智能水平领域风险治理的主导逻辑。本文从"技术双向性"的角度,围绕技术赋能和社会嵌入两个关键特征,结合人工智能技术发展和社会风险形成的历时性过程,构建了基于技术治理和制度治理协同演化的试探性治理框架。以美国、欧盟和中国的人工智能深度伪造技术治理为典型案例,进一步阐明了上述概念框架的互动机制、平衡关系与演化机理。研究得到的结论如下。

一是技术治理由"被动检测"逐渐发展到"主动检测"。人工智能技术赋能社会千行百业,具有极强的溢出带动性。研究发现,在技术治理层面,与科技界、学术界的普遍担忧不同,美国、欧盟和中国基本上都采取了鼓励创新和防范风险兼顾的谦抑性治理思路,同时随着知识的积累和风险事件的出现,不断及时、灵活地调整技术治理的强度。深度伪造技术治理的案例显示,在该技术兴起的早期阶段,技术治理的聚集点主要集中在伪造特征识别与提取等方面的"被动检测",而这种"发现才处理"的防御性治理方式,在某种意义上为深度伪造技术进一步发展预留了空间。当深度伪造技术逐步成熟,各种风险和危害事件频发,特别是在制度治理方式发生重大调整之后,技术治理的聚集点则由伪造特征识别与提取的"被动检测",转向为基于深度学习的特征鉴别和基于融合边界的伪造检测等"主动检测"技术,而这种"主动发现"的进攻性治理方式,在一定程度上阻断了深度伪造技术拓展边界的路径。

二是制度治理从"间接监管"逐渐转向"直接监管"。科技创新不断重塑社会,社会则不断适应、管理和重新定向科技创新。制度作为引导性和约束性的社会机制,在人工智能风险治理中发挥着基础性的作用。由于各国社会治理体制的差异,人工智能风险制度治理的主导方式和路径略有不同。美国表现为以权益保护为主导的上下分治型制度治理,欧盟表现为以行动框架为主导的集中型制度治理,中国表现为以场景应用为主导的部委协作型制度治理。虽然存在上述差异,但制度治理方式均表现出由"间接监管"到"直接监管"发展的特征。深度伪造技术风险治理的案例显示,在此技术发展的早期阶段,美国、欧盟和中国都较少颁布覆盖性的规制文件,多是针对风险

的聚焦点将深度伪造技术纳入数据信息监管、个人隐私保护等常规、宽泛的治理框架之中进行间接监管。当深度伪造技术的风险和危害进一步显现之后,美国、欧盟和中国都相继颁布了一系列直接监管的框架性文件,明确提出技术开发和使用的原则,具体规定开发者、应用者和管理者的主体责任和惩罚措施。

三是技制迭代发展由“治标型协调”演化形成“治本型协同”。试探性治理尝试通过谨慎、初步、灵活和渐进的方式把不确定性问题逐步转化为确定性问题,进而实现对鼓励创新和风险防范“二难困境”问题的辨证治理。研究发现,人工智能技术的试探性治理是通过技术治理和制度治理的协同演化过程实现的。在人工智能技术风险初露端倪的早期阶段,治理方式首先是“治标型协调”,即技术上防御性的“被动检测”和制度上谦抑性的“间接监管”。深度伪造技术治理的案例显示,在技术发展的早期阶段,各国先是尝试开发更加精准的深度伪造特征识别与提取技术,同时在制度上也作出“适当”调整,比如把深度伪造技术纳入生物信息保护、侵权责任处置等成熟的治理框架进行间接监管。随着深度伪造技术逐渐进入到政治、经济和社会等各领域,并产生一系列消极影响之后,治理方式则转变为“治本型协同”,即制度上扩张性的“直接监管”和技术上进攻性的“主动检测”。具体表现是直接颁布一系列的《法案》《指南》《指引》《规定》等框架性文件,收紧对深度伪造技术开发与应用的限制。同时,技术治理也由早期主要聚焦伪造特征识别与提取的被动检测技术,迅速调整为开发基于深度学习的特征鉴别和基于融合边界的伪造检测等主动检测技术。概言之,人工智能技术的试探性治理是在历时过程中,通过技术赋能社会和社会嵌入技术的迭代发展,实现了技术治理和制度治理的协同演化。

(二)进一步的讨论

本文把在治理实践中普遍存在,但在既有研究中未被充分重视的技制迭代协同演化问题,引入了人工智能风险治理的研究之中。通过将历时性视角和辨证思考带入试探性治理框架,生动还原了面对具有“乌卡特征”的治理难题时,技术治理和制度治理如何通过谨慎、初步、灵活和渐进的方式迭代发展,最终实现技术与制度协同演化的动态过程。尽管本文是以人工智能深度伪造技术治理作为分析案例,但在其他类似的新兴技术发展过程中,治理场域亦可能具有同样的结构因素和行动逻辑,试探性治理同样可以在技术赋能社会和社会嵌入技术的互动中,通过技制迭代协同演化的动态方式,从早期的“治标不治本”最终达到“标本兼治”的目的。

需要强调的是,试探性治理主要是一种适用于不确定性风险的治理模式,而且是在风险发展的不确定性阶段尤为适用。试探性治理并不排斥对确定性风险问题的传统治理。在本文的案例中,当深度伪造技术的风险和危害逐步显现,也就是不确定性风险转化为确定性风险之后,各国的治理模式旋即由谦抑性的“试探”迅速切换为扩张性的“监管”。具言之,面对由新兴技术引发的新兴治理难题,首要的问题不仅是选择一种适恰的治理模式,更为重要的是要树立一种弹性的思维逻辑,及时地根据事物的发展变化,因势利导、因时制宜地配置技术工具和制度工具。

本文研究也存在如下局限。一是,为了保证问题分析的深度,本文选择了案例研究方法,但在不同的技术分类领域中,技术属性、应用场景和治理场域均存在差异。因此,在试探性治理过程中,技制迭代协同演化规律可能不尽相同。二是,技制迭代协同演化也绝非必然,如因不同社会风险文化的包容性差异,某一技术的潜在风险在一种文化中可能难以忍受,而在另一种文化中也可能被高度宽容,本文虽对这一问题有所涉及但未展开深入的类型学分析,亦是研究的未尽之处。

参考文献:

[1]习近平. 加强领导做好规划明确任务夯实基础,推动我国新一代人工智能健康发展[N]. 人民日报, 2018-11-01(1).

[2]瞿晶晶, 王迎春, 赵延东. 人工智能社会实验:伦理规范与运行机制[J]. 中国软科学, 2022(11):74-82.

[3]庞祯敬, 薛澜, 梁正. 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, 2022,43(9):3-18.
[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 (3):108-123.
[5]贝克, 王武龙. 从工业社会到风险社会(上篇):关于人类生存、社会结构和生态启蒙等问题的思考[J]. 马克思主义与现实, 2003(3):26-45.
[6]吉登斯. 现代性的后果[M]. 田禾, 译. 北京: 译林出版社, 2011:31.
[7]贝克. 风险社会: 新的现代性之路[M]. 何博闻,张文杰,译. 北京: 译林出版社, 2004:36.
[8] KUHLMANN S, STEGMAIER P, KONRAD K. The tentative governance of emerging science and technology: a conceptual introduction[J]. Research policy,2019,48(5):1091-1097.
[9]杨伟, 周青, 方刚. 产业创新生态系统数字转型的试探性治理:概念框架与案例解释[J]. 研究与发展管理, 2020,32 (6):13-25.
[10]黄振辉. 多案例与单案例研究的差异与进路安排:理论探讨与实例分析[J]. 管理案例研究与评论, 2010(4):183-188.
[11]EISENHARDT K M, GRAEBNER M E. Theory building from cases: opportunities andchallenges[J]. Academic of management journal, 2007,50(1):25-32.
[12] COLE S. We are truly fucked: everyone is making AI generated fake porn now[EB/OL]. [2019-06-14][2023-12-10]. https://www. vice. com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley.
[13]程显毅, 谢璐, 朱建新, 等. 生成对抗网络 GAN 综述[J]. 计算机科学, 2019, 46(3):74-81.
[14]王振波, 吴湘玲. 数字时代"深度伪造"技术研究:机理特征、功能异化及其优化理路[J]. 北京航空航天大学学报(社会科学版),2022(5):1-9.
[15]波兹曼. 技术垄断: 文化向技术投降[M]. 何道宽, 译. 北京: 北京大学出版社,2007:18.
[16]赖志茂, 章云, 李东. 基于 Transformer 的人脸深度伪造检测技术综述[J]. 广东工业大学学报, 2023,40(6):155-167.
[17] ZHAO H Q, WEI T Y, ZHOU W B, et al. Multi-attentional deepfake detection [C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021:2185-2194.
[18]人工智能安全平台 RealSafe:瑞莱智慧 RealAI (real-ai. cn)[EB/OL]. (2023-04-14)[2023-12-10]. https://www. real-ai. cn/products/55. html.
[19] WANG J K, WU Z X, OUYANG W H, et al. M2TR: Multi-modal multi-scale transformers for deepfake detection [C]// Proceedings of the 2022 International Conference on Mu-ltimedia Retrieval. Newark, USA: ACM, 2022:615-623.
[20] AGARWAL S, FARID H, GU Y, et al. Protecting world leaders against deep fakes [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019:38-45.
[21] VAN DE WEGHE T. Six lessons from my deepfakes research at Stanford: how shou-ld journalists address the growing problem of synthetic media[EB/OL]. (2019-05-29)[2023-12-10]. https://medium. com/@ tomvandeweghe/six-lessons-from-my-deepfake-research-at-stanford-1666594a8e50.
[22]深度伪造内容检测平台 DeepReal:瑞莱智慧 RealAI (real-ai. cn)[EB/OL]. (2023-04-14)[2023-12-10]. https://www. real-ai. cn/products/9. html.
[23] Deepfake legislation: a nationwide survey—State and federal lawmakers consider legislation to regulate Manipulated media[EB/OL]. (2019-09-25)[2019-11-19]. https://www. jdsupra. com/legalnews/deepfake legislation-a-nationwide-86809/.
[24]张涛. 后真相时代深度伪造的法律风险及其规制[J]. 电子政务,2020(4):91-101.
[25] ROBERTSON A. Virginia's "revenge porn" laws now officially cover deepfakes[EB/OL]. (2019-07-01)[2022-10-17]. https://www. theverge. com/2019/7/1/20677800/virginia-revenge-porn-deepfakes-nonconsensual-photos-videos-ban-goes-into-effect.
[26]General data protection regulation[EB/OL]. (2018-05-25)[2019-11-19]. https://ec. europa. eu/commission/priorities/justice-and-fundamental-rights/data protection/2018-re form-eu-data-protection-rules/eu-data-protection-rules_en.
[27]European Commission. EU code of practice on disinformation [EB/OL]. [2019-09-22][2023-12-10]. https://www. hadopi. fr/sites/default/files/sites/default/files/ckeditor_files/1CodeofPracticeonDisinformation. pdf.
[28]董青岭. 数字外交中的深度伪造研究[J]. 中国信息安全, 2022(10): 66-69.

(本文责编:润　泽)