

人工智能伦理：规制、理论与实践难题

——学术史梳理及其问题域考察

岳璠

(东南大学 哲学与科学系, 江苏 南京 211189)

[摘要] 人工智能技术(AI)智能体面临包括决策权、所有权、隐私权等方面的伦理风险问题,也面临作为智能体在独立学习内容和独立行为目的等“类主体行为”中伦理能力方面的风险问题。为了规范发展,避免伦理风险发生,各机构、组织和企业建立了各自的准则和标准。国外学者关于AI伦理是否可行的探究,涉及AI主体形态(作为agent的)伦理复杂性问题、AI行为形态、AI位格形态关于道德权利(地位)、道德责任的争论。国内相关研究路径主要有三条:对AI的主体地位进行形而上学的分析、对AI伦理问题提出相应的实践解决框架、对AI的具体应用领域中的伦理问题进行探讨。从风险规制、理论拓展、实践难题三个方面对人工智能伦理进行学术史梳理及问题域考察,需要聚焦三大前沿:(1)推进系统化的AI伦理理论和原则框架;(2)建立“理论-实践”“全球性-地方性”相沟通的开放性解题思路;(3)形成探索AI道德形态研究的新方法。

[关键词] 人工智能伦理 理论-实践 道德形态

人工智能技术(Artificial Intelligence,以下简称AI)愈来愈“智能”并逐渐进入人们日常生活。AI智能体作为交互式、自主性、自学习的“自主体”,被称作“agent”(也可译做“类主体”或“准主体”)。它面临包括决策权、所有权、隐私权等方面的伦理风险问题,也面临作为智能体在独立学习内容和独立行为目的等“类主体行为”中伦理能力方面的风险问题。AI相关伦理问题与委托和责任等“代理模式”的深度运用及其智能化展现相关。现如今,这种模式的广泛应用,已经前所未有地将人工智能的伦理风险防范问题与保护人类生存安全和自由意志紧密地关联在一起。因此,AI伦理框架和伦理规制成为近几年国内外各方面关注的前沿和热点问题,也激发了伦理理论及其实践难题的持续研讨。本文从风险规制、理论拓展、实践难题三个方面,对AI伦理进行学术史梳理及问题域考察,试图借此思考AI伦理研究进一步发展与突破的方向。

一、问题的提出:AI伦理指南与风险规制

2004年国际首届机器人伦理研讨会正式提出“机器人伦理学”,机器行为模式被纳入伦理视域,AI伦理研究开始为世人所关注。

2019年5月22日,国际经合组织(简称OECD)理事会审议通过了经合组织数字经济政策委员会提出的“人工智能标准草案”。这是迄今为止世界上第一个政府间的AI标准。经合组织AI标准旨在通过加强对可信任AI的责任管理,确保AI尊重人权和尊重民主价值观,从而不断促进AI的技术创新并获得人们的信任。它的实体部分由两个部分组成:一是规定了可信任AI的责任管理,提出了五项基本原则(包容性增长、可持续发展和福祉;以人为本的价值观和公平性;透明性和可解释性;稳健性、可靠性和安全性;可追责性);二是规定了可信任AI的国家政策和国际合作^①。

[基金项目] 本文系国家社科基金重大项目“人工智能伦理风险防范研究”(20&ZD040)、江苏省社科基金重点项目“人工智能的伦理风险及防范研究”(20ZXA001)阶段性成果。

[作者简介] 岳璠,东南大学哲学与科学系教授,博士生导师,国家社科基金重大项目“人工智能伦理风险防范研究”首席专家,研究方向:人工智能伦理、科技伦理。

^① 唐川:《OECD制定人工智能发展建议》,《科研信息化技术与应用》2019年第3期。

为了规范发展,避免伦理风险发生,或在伦理风险发生后能够有解决机制,各种机构、组织和企业建立了各自的准则和标准,如:欧盟发布的《可信人工智能伦理指南草案》(Draft Ethics Guidelines for Trustworthy AI,2018)、《算法责任与透明治理框架》(A governance framework for algorithmic accountability and transparency,2019)以及《人工智能白皮书:通往卓越和信任的欧洲路径》(White Paper on Artificial Intelligence: A European approach to excellence and trust,2020);英国的《英国人工智能发展的计划、能力与志向》(AI in the UK: ready, willing and able,2018);生命未来研究所(Future of Life Institute)的阿西洛马人工智能原则(Asilomar AI Principles,2017);OpenAI的OpenAI宪章(OpenAI Charter,2018);谷歌发布的人工智能原则(Google AI Principles,2018);电气电子工程师协会(IEEE)的人工智能设计伦理准则(Ethically Aligned Design V2,2017);人工智能联盟(Partnership on AI)的人工智能联盟信条(Partnership on AI Tenets,2016);国际电信联盟(ITU)等联合国机构的人工智能造福人类峰会(AI for Global Good Summit,2017);哈佛大学、麻省理工学院(MIT)的人工智能伦理与监管基金会(Ethics and Governance of AI Fund,2017);德国自动驾驶伦理委员会(German Ethics Commission on Automated and Connected Driving)提出首套针对无人驾驶汽车的官方伦理指导原则(2018);欧洲政治战略中心发布的《人工智能时代:确立以人为本的欧洲战略》(The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines,2018);斯坦福大学创建“以人为本人工智能研究院”(Stanford Human-Centered AI Institute,2018);奥迪、百度、宝马等国际企业联合发布的《自动驾驶安全第一》白皮书(Safety First for Automated Driving,2019);腾讯公司开展的“科技向善”专题活动,以及马化腾提出的可知、可控、可用和可靠四个原则(2018);2024年3月21日,联合国大会通过首个关于人工智能的全球决议《抓住安全、可靠和值得信赖的人工智能系统带来的机遇,促进可持续发展》。

中国国务院于2017年7月20日发布《国务院关于印发新一代人工智能发展规划的通知》,作为我国人工智能发展规划的基石^①,并前后出台《“互联网+”人工智能三年行动实施方案》^②、《新一代人工智能治理原则——发展负责任的人工智能》^③、《北京市自动驾驶车辆道路测试管理实施细则(试行)》^④;2020年7月27日,国家标准化管理委员会等五部门印发了《国家新一代人工智能标准体系建设指南》,明确指出伦理、安全、隐私在AI发展中的引领作用,规范AI服务冲击传统道德伦理和法律秩序而产生的要求,重点研究医疗、交通、应急救援等特殊领域的AI伦理问题^⑤;2021年9月25日,国家新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》,旨在将伦理道德融入人工智能全生命周期,积极引导全社会责任的开展人工智能研发与应用活动^⑥;2022年4月中国信息通信研究院发布《人工智能白皮书》指出,人工智能接下来的持续健康发展,将由“技术创新、工程实践、可信安全‘三维’坐标来定义和牵引”,人工智能治理实质化进程加速推进,从初期构建以“软法”为导向的社会规范体系,开始迈向以“硬法”为保障的风险防控体系,特别聚集于自动驾驶、智慧医疗和人脸识别等领域^⑦;2023年3月国家人工智能标准化总体组、全国信标委人工智能分委会发布《人工智能伦理治理标准化指南》,其中明确提出以人为本(For Human)、可持续性(Sustainability)、合作(Collaboration)、隐私(Privacy)、公平(Fairness)、共享(Share)、外部安全

① 《国务院关于印发新一代人工智能发展规划的通知》(2017-07-20)[2024-01-10],https://www.gov.cn/zhengce/zhengceku/2017-07/20/content_5211996.htm。

② 《“互联网+”人工智能三年行动实施方案》(2016-05-23)[2024-02-13],<https://www.gov.cn/xinwen/2016-05/23/5075944/files/9cb49ac44cf341b29adf687b6857da34.pdf>。

③ 《新一代人工智能治理原则——发展负责任的人工智能》(2019-06-17)[2024-01-27],https://www.gov.cn/xinwen/2019-06/17/content_5401006.htm。

④ 《北京市自动驾驶车辆道路测试管理实施细则(试行)》(2020-11-12)[2024-01-27],https://www.beijing.gov.cn/xxgk/flfg/zcfg/202011/t20201116_2136351.html。

⑤ 《国家新一代人工智能标准体系建设指南》(2020-07-27)[2024-02-17],https://www.gov.cn/zhengce/zhengceku/2020-08/09/content_5533454.htm。

⑥ 《新一代人工智能伦理规范》(2021-09-25)[2024-03-09],https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html。

⑦ 中国信通院:《人工智能白皮书(2022年)》(2022-04-03)[2024-02-17],https://www.thepaper.cn/newsDetail_forward_17646199。

(Security)、内部安全(Safety)、透明(Transparency)、可问责(Accountability) 10 项可实施性较强的人工智能伦理准则^①。

二、AI 伦理的学术史展开：理论与实践的复杂性

在各国和各组织制定相应政策框架的同时,学术界关于 AI 伦理问题进行了激烈的思考和讨论。国外学者最早切入 AI 伦理是否可行的问题探究,揭示了 AI 伦理在其学术史视域显现的理论与实践的复杂性,主要涉及三个方面: AI 主体形态(作为 agent)的伦理复杂性问题; AI 行为形态的伦理难题; AI 位格形态关于道德权利(地位)、道德责任的争论。

第一, AI 主体形态的伦理复杂性问题。

温德尔·瓦拉赫(Wendell Wallach)认为,如果把重点放在实践上,是过分简化了伦理问题的复杂性。伦理复杂性有两个根源:一是伦理理论内部关于基本概念的争论;二是对现实世界做规范化判断的分歧。人类道德是一种复杂的活动,期望 AI 智能体立即解决所有问题是不切实际的,但我们应该持有一种开放性立场:任何能够提高机器人道德考量的敏感性的进步,无论多么微小,都是在朝着正确的方向迈进^②。雅恩·乐昆(Yann LeCun)认为,人们对 AI 最大谬见有二:一是“AI 不会有情感”,事实上它们很可能会有情感;二是“如果 AI 有情感,它们将会和人类情感一模一样”,事实上 AI 没有理由拥有自我保护直觉、嫉妒等情感。不过,我们可以将“利他主义”或其他对人类有利的情感注入到 AI 里面,让 AI 能够取悦人类,让人类与之交互,融入人类生活。未来,绝大多数 AI 将会变得更加专业化,但是却不会有情感。你的汽车自动驾驶装置只会为你开车,不会和你谈恋爱^③。黛博拉·约翰逊(Deborah Johnson)是标准派的代表,他认为物体要能够成为道德主体,必须满足清晰明确的条件,其中尤其强调物体的心理因素如自由意志、意识、欲望等^④。卢西亚诺·弗洛里迪和 J. W. 桑德斯(Luciano Floridi & J. W. Sanders)反对对于道德主体的标准论看法,认为道德主体并不必须表现出自由意志、精神状态和责任,而应该从“无意识道德”(mind-less morality)角度思考,他们强调人工智能体的交互性(与环境互动)、独立性(有能力改变自身和它的相互作用)和适应性(可能会基于与环境互动的结果改变自身及其在相互作用中的实现方式)。这不仅对网络空间有效,而且对于生态系统、动物等都是适用的。人工智能体和人类之间的关系类似于儿童和成年人的关系,人工智能体尚未获得完整的道德地位^⑤。菲利普·布雷(Philip Brey)反对将人工物看作完整的道德主体,认为这样的观点模糊了人的能动性和人造物的能动性之间的重要区别,以及抹杀了人的行为的独特性^⑥。琳达·约翰逊、弗朗西丝·S. 格劳森斯基(Linda Johnson & Frances S. Grodzinsky)认为人工智能体的道德地位是人类主体地位的分有,不是机器自己有自由意志,而是机器行动的“意向性”体现了人类设计师的意向性,并能产生善或恶的道德后果,因此具有一定的道德性^⑦。库科尔伯格(Mark Coeckelbergh)在《日趋发展的道德关系》中认为,道德地位是在主体之间和主客体之间的关系之中呈现出来的^⑧。摩尔(James Moor)的观点更为细致和中立,他在其《Four Kinds of Ethical Robots》一文中区分了四种人工道德主体——受影响的人工道德主体(ethical impact agents)、内隐的人工道德主体(implicit ethical agents)、外显的人工道德主体(explicit ethical agents)和完全的人工道德主体(full ethical agents),认为人工智能在不同的发展阶段应该被看作不同的道德主体,并对四种道德主体之

① 国家人工智能标准化总体组、全国信标委人工智能分委会:《人工智能伦理治理标准化指南(2023年)》(2023-03-13)[2024-02-10], https://baike.baidu.com/item/人工智能伦理治理标准化指南/62961103?fr=ge_alas。

② Wendell Wallach, Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxford University Press, 2009.

③ 《对于人工智能,你或许理解错了》(2016-06-21)[2024-01-21], <http://www.gongkong.com/article/201606/67442.html>。

④ Deborah G. Johnson, “Computer Systems: Moral Entities But Not Moral Agents”, *Ethics and Information Technology*, 2006(4).

⑤ Luciano Floridi, Sanders, J. W., “On the Morality of Artificial Agents”, *Minds and Machine*, 2004, 14(3), p. 349 - 379.

⑥ Brey P., “From Moral Agents to Moral Factors: The Structural Ethics Approach”, *Moral Status of Technical Artifacts*, 2014(17), pp. 125-142.

⑦ Grodzinsky, F.S., Miller, K.W., Wolf, M.J., “The Ethics of Designing Artificial Agents”, *Ethics and Information Technology*, 2008(3), 10(2-3), pp. 115-121.

⑧ Coeckelbergh M., *Growing Moral Relations: Critique of Moral Status Ascription*, New York: Palgrave Macmillan, 2013.

间进行了较为详细的区分,建议将外显的 AI 作为机器伦理的范例目标,这样既可以保证其生活中的复杂性和实践中的重要性,又不会成为人类生存和伦理的威胁^①。

第二,在 AI 行为形态上,关于 AI 进行道德推理和道德决策的研究。

从技术的可实现角度出发,道德机器人的伦理植入路径有自上而下的伦理原则植入、自下而上的伦理学习和混合式进路。其中自上而下的伦理原则植入是基于规则推理而自下而上学习则基于学习算法、基于脑认知结构、基于马尔可夫决策和基于决策函数。从基于规则推理路径出发的研究包括:Bringsjord S. 等人通过分析机器人三原则的不足,总结了两种基本机器决策原则,一是机器只做允许的动作,二是机器在做强制性动作时会受到其他可用动作的制约^②。Anderson 夫妇等根据生物学伦理原则,设计出了名为 MedEthEx 的伦理决策系统,该机器能够处理一定的护理人遇到的伦理难题。他们还以功利主义和义务论为原则开发了两种伦理决策系统,提出一种可以确保自主系统伦理行为的 Case-Supported Principle-based Behavior(CPB) 范式。该范式的实现是由伦理学家在实际伦理案例中得出统一的伦理原则,进而自主系统在该原则的指导下作出决策。该方法保证了原则的准确性并验证了案例的可解释性^③。此外,McLaren、Arkin、Wynsberghe 也提出了以不同伦理原则为基础的道德决策机器。从基于学习算法路径出发的研究,包括 Armstrong S 等使用贝叶斯理论构建一种通过效用函数来进行智能体的决策设计模型。Briggs G、Clore G L、Franklin S 等从基于脑认知结构出发进行了探讨,Coeckelbergh M、Rysewyk、Rzepka V 等则从混合路径上提出了相应的理论模型^④。

不同的伦理实现路径有不同的优缺点。从 AI 的本质和伦理角度出发,采用何种实现路径更为恰当需要理论和实践相结合进行选择。安德森夫妇已经通过自上而下的方式将尊重自主权、行善和不伤害的原则嵌入了医疗伦理专家系统 MedEthEx 之中。出于技术和对伦理认知的考量,许多哲学家都认为自上而下的道德植入是必不可少的,因为如果人没有言行一致的普遍应用的道德准则,行为便不可能做到理性。机器也一样,只有在一定的理论框架之上,才能更好地理解道德本身。瓦拉赫、艾伦、丘奇兰德等提倡用自上而下理论植入、结合联结主义学习的混合进路,让机器能够应对更加复杂的情况。从目前的学界和应用中,混合的伦理实现路径是一种较为普遍认可的方法,混合路径既采用了自上而下的伦理原则植入,也采用了自下而上的规则学习,更像是人的学习进步过程,既能满足对普遍规则的遵守,也能适应于复杂多变的现实环境,是未来 AI 伦理实现的重要研究方向。埃吕尔(Ellul J.) 认为,目前人们对于“效率”的普遍崇拜和将其作为目的而远离了人文主义,走向了技术主义,形成了“力量伦理”,使得技术成为社会的主导者,但是力量的主导和发展的最终结果却是技术对人类的统治和对人类自由的剥夺。因此,为了人类的自由,应该建立起以自由(freedom)、非力量(non-power)、包容冲突(inclusive conflict)和鼓励越界(encourage transgression)为基础的“非力量伦理”,尝试建立一种对技术的新的认识和观念^⑤。Cathy O'Neil 在《Weapons of Math Destruction: How big data increases Inequality and Threatens Democracy》一书中提出大数据并不是人们所认为的对人类有那么大的益处,相反,除了在找到处于困难中的人时是能够提供较大帮助,在其他方面比如惩罚他人等并不具有任何好处,反而会在权利和偏见的影响下加强社会的不平等性和威胁民主^⑥。

① Moor J H., “Four Kinds of Ethical Robots”. *Philosophy Now*, 2009(72), pp. 12-14.

② Bringsjord, S., Arkoudas, K., Bello, P., “Toward a General Logicist Methodology for Engineering Ethically Correct Robots”, *IEEE Intelligent Systems* (2006), DOI: 10.1109/MIS.2006.82.

③ Anderson, M., Anderson, S. L., “Armen, C. MedEthEx: A Prototype Medical Ethics Advisor”, *Proceedings of the 18th conference on Innovative applications of artificial intelligence*, Volume 2 July 2006, p. 1759 - 1765. Anderson, M., Anderson, S. L., “Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm”, *Industrial Robot*, 2015, 42(4), pp. 324-331.

④ Coeckelbergh M., *Growing Moral Relations: Critique of Moral Status Ascription*, New York: Palgrave Macmillan, 2013.

⑤ Ellul J., *The Power of Technique and the Ethics of Non-Power*, LI Lun, YX Pan, “From Power Ethics to Non-Power Ethics: Jacques Ellul's Theory about Freedom in the Technological Society”, *Studies in Dialectics of Nature*, 2018(11), pp. 33-38.

⑥ Cathy O'Neil., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Press, 2016. ISBN: 9780553418811.

第三, AI 位格形态关于道德权利、责任问题的争论。

价值中立论学者反对将人工物纳入道德主体的范畴,并反对人工物的道德属性,认为只有人是具有道德属性的,人工物的存在不具位格性(Person),只是一种服务于人的工具性存在,并不具有道德上的属性。如约瑟夫·C.皮特(Pitt J. C.)认为技术人工物不包含、不具有、也不展现价值,人工物的价值正是通过人类的决策过程产生的,其中所体现的很多种价值,其实都是源自人本身的价值^①。皮特·克罗斯(Kroes P.)对价值中立主义的观点进行考查,区别了其中的价值——内在价值和工具价值/关系价值/外在价值,内在价值是因自身目的而具备的价值,工具价值是作为人的工具使用而产生的对人的价值。对人工物而言,在不同阶段体现三种不同的价值:预期价值(设计者预期的价值)、体现价值和实现价值(在实际使用中实现的价值)^②。约翰逊(Johnson)也认为计算机不是道德主体,而只能作为道德实体,因为计算机不能满足康德等传统伦理观中所认为的那样完全符合道德能动性的基本标准,道德主体的关键在于意向性行为,因为意向性是主体自由的体现^③。拉图尔(Latour B.)从后现象学角度反对人类中心主义的个体主义伦理学。他认为人和非人的存在是彼此依存的,相互作用的,是意向性不可或缺的两个部分。人工物是以类似于人类的方式在执行道德规则,因此,人类和人工物都是社会规则的承担者,人不是唯一的道德主体,人工物也是道德主体^④,因为单独的人是无法完成一件事的。在减速带例子中,就包含了减速带、设计师、司机甚至是整个交通系统,因此,道德主体既不是司机也不是设计者和减速带,而是各种实体的集合。拉图尔、鲍尔斯和约翰逊认为,人和人造物等共同作用进行了道德行为,不过也都承认人工物不能脱离与人类的存在而成为道德主体。不同的是,鲍尔斯和约翰逊承认在人与人工物之间存在不对称性,即人工物不能脱离人而完成道德行为,而人可以自主进行道德行为。相反,拉图尔认为不仅仅人造物不能脱离人完成道德行为,人也不能脱离人造物而进行道德行为,就像没有人,汽车的安全带便无意义,而乘客系安全带也不仅仅是自己意志的结果,而是车上的提示灯、设计师的意图、交通警察和交规等共同引起的。因此,人的能动性并不是人的主体性造成的,而是人、人造物等实体所组成的多元网络的一种属性,在这个网络中,多个主体共同作用产生一个特定的行为,人和人造物在本体论上是对称的^⑤。艾安娜·霍华德(Ayanna Howard)认为:学习算法已经普遍运用,但种族歧视、性别歧视也被算法学习进去,尽管算法很聪明,但他们仍然保持着每个社会相同的一些偏见。他们在数据集中发现并反映这种内隐的偏见,并以此强调和强化这些偏见,并认为这就是全球真理。她认为偏见融入当前 AI 系统的具体例子,以及偏见会影响未来此类系统的设计。詹森·鲍任斯坦(Jason Borenstein)提出人类应该允许同伴型 AI“推动”他们的人类用户朝着“更合乎道德”的方向发展,这需要以罗尔斯正义原则来说明 AI 如何在人类身上培养“社会公正”的倾向^⑥。约翰·P. 萨林斯(Sullins J. P.)认为:如果 AI 有抽象层次的自主意图和责任感时我们就应该认为 AI 是道德代理体,如果 AI 可以从很多方面被看做是自主的,这个 AI 就是智能道德体,有可能接近甚至超越人类的道德形态。因此,如果我们追求这项技术,未来高度复杂的交互式 AI 将是具有相应权利和责任的道德体,今天 AI 可以部分被看作一种非抽象级别的道德体^⑦。

三、AI 伦理的基础问题域:形上根据与实践框架

相比于国外学者的研究而言,国内学术界对 AI 伦理研究起步较晚,主要从 2017 年开始兴起。

① 约瑟夫·C.皮特:《技术思考》,马会端、陈凡译,沈阳:辽宁出版社,2012年。

② Kroes P., Verbeek P. P., *The Moral Status of Technical Artefacts*, Dordrecht: Springer Press, 2014. ISBN: 9789400779143.

③ Johnson D. G., "Computer System: Moral Entities but not Moral Agents", *Ethics and Information Technology*, 2006, 8(4), pp. 195-204.

④ Latour B., "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts", In W. E. Bijker & J. Law (Eds.), *Shaping Technology—Building Society: Studies in Sociotechnical Change*, Cambridge: MIT Press, 1992, pp. 225-258.

⑤ Latour B., *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford: Oxford University Press, 2005.

⑥ Ayanna Howard, Jason Borenstein, "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity", *Science and Engineering Ethics*, 2017. DOI: 10. 1007/s11948-017-9975-2.

⑦ Sullins J. P., "When is a Robot a Moral Agent?", *International Review of Information Ethics*, 2006(6).

国内的相关研究路径主要有三条:对 AI 的主体地位进行形而上学的分析、对 AI 伦理问题提出相应的实践解决框架、对 AI 的具体应用领域中的伦理问题进行探讨。

第一,对 AI 的主体地位进行形而上学的分析。

传统的社会伦理一向是以调节人与人的关系为中心,兼论人与物的关系。刘大椿认为智能革命将成为人类未来的大趋势,智能革命给所有人提出了人性本质问题:无论智能革命如何推进,人类生命和社会文明如何演化发展,怎样使每个人在由人类所开创的深度智能化的未来都有事情可做?如何让一切人群都在心理和精神上呈现出朝气蓬勃和幸福快乐的状态,而不是自惭形秽、百无聊赖的失败者?^①何怀宏提出未来的伦理学大概还需要将人与智能机器的关系纳入其考虑范围。这主要是因为人赖以获得对其他所有自然物的支配优势的智能,将可能被机器超越。因此,关于人机关系的伦理思考,何怀宏提倡底线思维,即抑恶应该优先于扬善,放慢技术的发现速度,优先和集中地考虑规范智能机器的手段和限制其能力,而不是考虑如何设定和培养机器对人类友好的价值判断,让机器保持在“物”的水平,一切对人类自由和身心安全有关的都必须由人来完成^②。王天恩认为人工智能发展呈现出的学科一体化趋势,空前凸显了马克思主义理论的整体层次优势。在马克思主义人类解放学说的整体关照中,可以对人工智能发展有一个更高层次的把握。人工智能的发展晓示了人类解放的现实可能性,并将具体通向真正的人类历史。由人类解放的人工智能发展前景,可以越来越清晰地窥见:人工智能之“镜”为人的自由全面发展提供了前所未有的参照,人工智能的信息展开为人的类解放提供了深刻启示^③。陈凡、程海东提出人工智能的发展和应引发了人们对于“奇点”的关注。实质上,人工智能还处于发展的早期,只能对人类智能进行模拟,它的产生和发展是为了人类的生存和发展,因而与其他技术一样,是人类与自然交流的方式之一。而且与人类智能相比,人工智能不具备自我意识,也无从把握人类的意义,更不能独立从事人类的实践活动。因此,在可预见的未来,人工智能并不会独立获得主体性,也不会成为超越人类的存在^④。赵汀阳从政治哲学的角度为人工智能的伦理研究进行了辩护。他认为人工智能的发展问题最终是个政治问题。人类首先需要一种世界宪法,以及运行世界宪法的世界政治体系,否则无法解决人类的集体理性问题。人类至今尚未发展出一种能够保证形成人类集体理性的政治制度,也就无法阻止疯狂的资本或者追求霸权的权力。在低技术水平的文明里,资本和权力不可能毁灭人类,但在高技术水平的文明里,资本和权力已经具备了毁灭人类的能力。更危险的是,资本和权力的操纵能力正在超过目前人类的政治能力,因此,要控制资本和权力,世界就需要一种新政治,即天下体系。天下体系的一个重要应用就是能够以世界权力去限制任何高风险的行为。因为一旦 AI 成为超级智能体,人类的生存就将受到巨大的威胁^⑤。

成素梅主编出版的《信息文明的哲学研究丛书》收入了《人工智能的哲学问题》(成素梅、张帆)、《信息文明的伦理基础》(段伟文)、《大数据时代的认知哲学革命》(戴潘)、《虚拟现象的哲学探索》(张怡)和《人的信息化与人类未来发展》(计海庆)五本专著,是国内在信息文明的哲学反思方面较系统的研究成果。成素梅、张帆认为信息文明演进的高阶是智能社会,哲学家应该提供一种前瞻性的研究,人工智能发展应以人为尺度,体现人的目的性,并融入人类文化之中。段伟文则提出,“我们需要有一个与其他价值系统对接的价值接口,这一现实需求将倒逼数字经济与智能社会建设展开进一步的价值建构”,目前有大量的数据伦理和人工智能伦理规范与准则,“接下来就是要通过对这些原则在不同价值系统下的内涵的理解与对话,提升价值上互操作性”^⑥。孙波和周雪健对人工智能则保持乐观的态度,认为我们不必对 AI 技术恐慌或技术崇拜,而应该从“功能偶发性失常”角度看待

① 刘大椿等:《智能革命与人类深度智能化前景(笔谈)》,《山东科技大学学报(社会科学版)》2019年第1期。

② 何怀宏:《人物、人际与人机关系》,《探索与争鸣》2018年第7期。

③ 王天恩:《人类解放的人工智能发展前景》,《马克思主义与现实》2020年第4期。

④ 陈凡、程海东:《人工智能的马克思主义审视》,《思想理论教育》2017年第11期。

⑤ 赵汀阳:《人工智能“革命”的“近忧”和“远虑”——一种伦理学和存在论的分析》,《哲学动态》2018年第4期。

⑥ 段伟文:《信息文明的伦理基础》,上海:上海人民出版社,2020年,第261页。

AI 并建立相应的伦理^①。徐英瑾认为 AI 研究需要多学科领域交叉进行,将伦理规范转化为机器编码在逻辑上必须预设一个好的语义学理论框架,而目前主流 AI 研究所采用的语义学基础理论却恰恰是成问题的。他主张引入认知语言学的理论成果,特别强调“身体图式”在伦理编码进程中所起到的基础性作用,主张机器伦理学的核心关涉将包括对 AI“身体”的设计规范问题^②。

孙伟平与戴益斌从对人类的主体地位分析出发,得出主体地位是由存在论、认识论和价值论三个角度综合界定的,而人工智能对应于这三点,都不符合主体的含义,因此人工智能并不具备主体地位。不过,AI 也不仅仅是简单的工具,有的 AI 可以被当做工具性的存在,有的 AI 应该被界定在工具与主体之间,即处于一种准主体地位。戴益斌还从责任角度出发,沿用汉森的观点,将 AI 看做一种“扩展主体”,并根据不同的 AI 产品设定不同的责任指派方案,明确不同情景下的联合责任^③。颜青山认为,AI 道德地位问题是一个真正新的道德难题,它可表述为“我们如何确定人工智能是或不是一个人格实体?”解决这个难题涉及“他心难题”及其变种“机心难题”。解决他心难题的方案对人工智能都可能是失效的。机心难题本质上是一个元伦理学或道德形而上学问题,但这个在元伦理学层次无法解决的问题可以在规范伦理学的层次上得到解决。他根据人格伦理学和手段伦理学的规则提出对待人工智能的原则——“选言命令式”,即应当尊重一个尊重你的实体的运行规则:当它是心灵实体时,尊重其本身;当它是一个非心灵实体时,尊重制造它的人,从而悬置了心灵本身,通过其行为进行判断^④。

第二,对 AI 伦理问题提出相应的实践解决框架。

由于 AI 发展的迅速性、应用的普遍性和影响的深入性,国内学者更多对 AI 带来的伦理、社会影响进行了研究,并提出了相应的解决框架。于雪、段伟文从人机关系出发,认为人工智能给人带来的主要伦理问题是其作为人与人之间的第三者所带来的,面对 AI 发展所产生和即将到来的大量伦理问题,应该遵循整体性、过程性、适应性、相容性、灵活性和鲁棒性等原则,从技术伦理角度对 AI 进行实质性分析,明确 AI 伦理的价值、标准和责任,进而建构以“技术-伦理”和信任一体的行业协作创新机制。通过社会各方协同建立人工智能伦理治理和体系框架,制定相应的风险防范机制,并确保人类和平的基本底线不被触碰^⑤。孙伟平对人工智能的积极社会效应提出了自己的看法,他认为人工智能具有积极的社会效应,但由于人们思想观念滞后,政策取向不清晰,伦理规制缺失,法律法规不健全,人工智能使人类面临巨大的不确定性和风险,造成诸多社会价值冲突和伦理困境,如挑战人的本质、价值和人类的道德权威,冲击传统的伦理关系,造成数字鸿沟,解构社会,而人工智能带来的伦理问题正好成为新时期道德伦理建设的“助推器”。我们必须对人工智能及其应用后果进行全方位的价值反思,制定智能社会的价值原则与综合对策,秉持人本原则、公正原则、公开透明原则、知情同意原则和责任原则,将人工智能纳入健康发展的轨道^⑥。李伦、孙保学将人工智能伦理研究方向分为四个板块,即人工智能道德哲学、人工智能道德算法、人工智能设计伦理和人工智能社会伦理,并从本体论角度分析了人工智能的自主性。他们建议从责任伦理角度处理人机关系,即站在人类角度和未来角度为 AI 的设计负责,并在此基础上提出了相应的责任分配原则:以人为本、共生共存原则;人类作为责任主体承担全部后果;分级分类制订担责方案。以及从人工智能体的“输入-输出”能否控制,在四种语境下给出了人机系统具体的责任分配方案^⑦。

王淑庆认为将逻辑学与伦理学结合的途径会给 AI 的研究带来新思路,欲使人工智能体具备道德决策能力,一种可设想的工作是基于形式伦理,即把伦理原则或规则形式化。不同于大众对于 AI 的伦理植入视角,他认为让 AI“知道”哪些行为符合道德很重要,更重要的是让其知道哪些行为违反

① 孙波、周雪健:《人工智能“伦理迷途”的回归与进路》,《自然辩证法研究》2020年第5期。

② 徐英瑾:《具身性、认知语言学与人工智能伦理学》,《上海师范大学学报(哲学社会科学版)》2017年第6期。

③ 孙伟平、戴益斌:《关于人工智能主体地位的哲学思考》,《社会科学战线》2018年第7期;戴益斌:《试论人工智能的伦理责任》,《上海大学学报(社会科学版)》,2020年第1期。

④ 颜青山:《对待人工智能的选言命令式》,《社会科学》2018年第12期。

⑤ 于雪、段伟文:《人工智能的伦理建构》,《理论探索》2019年第6期。

⑥ 孙伟平:《关于人工智能的价值反思》,《哲学研究》2017年第10期。

⑦ 李伦、孙保学:《给人工智能一颗“良心(良心)”》,《教学与研究》2018年第8期。

道德标准,真正“道德上值得称赞的人工智能体”(MPAAs)应该具有“有意不为”的能力,尤其是“忽略”和“抑制”的能力^①。张正清和黄晓伟认为,从责任的他者视角出发对智能机器道德自主性与意向性的质疑,并不能否认人工智能具有道德责任能力。从他者期望型道德责任的角度看,智能机器由于人们的期望偏见、智能偏见和地位偏见而不被承认具有道德责任能力。如果智能机器想要在现代技术社会的责任网络中获得应有的地位,就需要我们如实地期望智能机器人的道德责任,摒弃事后立场、语境不匹配、人类中心主义等经典责任伦理面临的问题,从而生成新的他者期望型道德能动者^②。李熙等讨论了人工智能的研究目的,认为机器学习为了获得通用性必须诉诸形而上的“善”,但仅有形而上的基本“善”远不能保证智能体的行为符合人类的主流价值观,为确保人类利益,还需要为机器赋予人类的价值观,最直接的方式是为机器赋予符合人类利益的效用函数,但智能体在计算期望效用最大化以追逐功利主义的“善”的过程中,不可避免地会侵占人类的资源。为了让机器符合人类利益,保留关机中断权,把风险降到最低限度,需要巧妙融合形而上的“善”与功利主义的“善”,进行“先验”与“效用”的转化,并灵活运用逆强化学习或价值强化学习^{③④}。

第三,对 AI 的具体应用领域中的伦理问题进行探讨

部分学者对人工智能在无人驾驶汽车、医护、伴侣机器人等具体应用领域的伦理问题进行了相关研究。陈齐平,魏佳成,钟陈志鹏,罗玉峰和王亮从技术角度对目前机器伦理决策的实现路径进行了分析和比较,并对无人驾驶汽车的伦理植入路径进行了分析,提出将人类道德规范应用到机器伦理决策设计、运用机器学习实现伦理决策设计、利用法律及行业标准规范来实现机器伦理决策设计三个重要研究方向,以为 AI 伦理植入和实现提供更好的基础^⑤。赵汀阳在《有轨电车的道德分叉》中认为电车难题并不是难题,相反,人们应该更多地关心现实生活,这有利于大家共在的善^⑥。李德顺在《价值独断主义的终结——从“电车难题”看桑德尔的公正论》中也对电车难题进行了消解^⑦。李醒民对人工智能发展可能引起的诸多伦理问题进行了探讨,包括失业问题、算法偏见问题、隐私问题、心理依赖问题、促使人懒惰问题和社会两极分化问题等,从总体上出发给出了人工智能发展相应的原则,即以人为本、预防风险、伦理建设、明确责任、知识普及和监管到位^⑧。刘程考察了 AI 时代的隐私伦理建构,认为作为一种自觉的伦理实践,它必须建立在一种批判性考察基础之上^⑨。另外,游辉辉、马永慧、黄立文等人对伴侣机器人的相关伦理问题进行了研究(《伴侣机器人应用的伦理、社会问题探讨》《伴侣机器人的道德问题研究》)^⑩;王健、林津如,郑婧萱对老年护理机器人带来的孝养关系疏远、孝道降阶化等伦理风险展开了研究^⑪。

综上所述,目前国内外相关学者及相关行业的一个基本共识是:“AI 将会在未来几十年对人类社会产生巨大影响,带来不可逆转的改变”,而“AI 伦理将是未来智能社会的发展基石”。AI 智能体对人类日常生活的影响是全面而深刻的,它将塑造新的交往方式、教育方式、医疗模式、饮食方式、出行方式、军事行为、政治经济行为等各个方面。已故著名科学家霍金预见到 AI 带来的风险,发出了“警惕人工智能”的警示。在诸多风险中,AI 智能体与人类相与的“共生型”(或“互生型”)的道德形态方面的风险,将是最为突出且最为紧要的伦理风险,许多科幻作家、怀疑论者、技术悲观主义者(包括一些温和的乐观论者)或多或少会将机器人列入奴役或取代甚至吞噬人类的“怪物”之行列。

① 王淑庆:《人工智能体“有意不为”的伦理意蕴》,《东北大学学报(社会科学版)》2020年第3期。

② 张正清、黄晓伟:《作为“他者”而承担道德责任的智能机器》,《道德与文明》2018年第4期。

③ 李熙、周日晴:《从三种伦理理论的视角看人工智能威胁问题及其对策》,《江汉大学学报(社会科学版)》2019年第1期。

④ 王银春:《人工智能的道德判断及其伦理建议》,《南京师大学报(社会科学版)》2018年第4期。

⑤ 陈齐平、魏佳成、钟陈志鹏、罗玉峰、王亮:《智能机器伦理决策设计研究综述》,《计算机科学与探索》2019年第11期。

⑥ 赵汀阳:《有轨电车的道德分叉》,《哲学研究》2015年第5期。

⑦ 李德顺:《价值独断主义的终结——从“电车难题”看桑德尔的公正论》,《哲学研究》2017年第2期。

⑧ 李醒民:《人工智能技术性科学与伦理》,《社会科学论坛》2019年第4期。

⑨ 刘程:《隐私辩护的伦理正当性及其可能进路》,《江海学刊》2023年第2期。

⑩ 游辉辉、马永慧:《伴侣机器人应用的伦理、社会问题探讨》,《中国医学伦理学》2019年第8期;黄立文:《伴侣机器人的道德问题研究》,《湖南科技学院学报》2019年第6期。

⑪ 王健、林津如:《护理机器人补位子女养老的伦理风险及其防范》,《道德与文明》2019年第3期;郑婧萱:《智能护理机器人的伦理风险防范研究》,东南大学研究生院,2022年。

四、AI 伦理研究的发展与突破

为了人类福祉,在 AI 的发展与应用中,伦理必须占据中心地位。这一条已成为国内外学者的基本共识。这条共识隐含的核心思想是:与 AI 发展相随的必是制定合适的伦理指导原则以防范其风险。这需要对 AI 的伦理风险的起因、影响、规则制定的方法以及效果评估等先行进行预备性前瞻思考,以利于 AI 伦理的相关标准建立和风险防范。

综合梳理国内外人工智能伦理的学术研究成果,可以发现:目前的重点是要突破政策制定者、技术专家、科学家、行业领导者和人文社会科学领域的专家(特别是伦理学家和工程技术专家)各说各话的弊端,亟须在以下三个方面发展和突破。

第一,亟需推进系统化的 AI 伦理理论和原则框架的研究。

尽管国外 AI 伦理研究早于国内且研究成果更丰富,但国内外研究均还不具备体系化的 AI 伦理的理论和原则框架。国外学者如迈克尔·安德森和苏珊·利·安德森(Machine Ethics)、迈克尔·R.W.道森(Minds and machines),荷兰学派维克克(Moralizing Technology Understanding and Designing the Morality of Things)、大卫·J.冈克尔(The Machine Question Critical Perspectives on AI, Robots, and Ethics)、瓦拉赫和艾伦(《道德机器——如何让机器人明辨是非》)等,出版了 AI 伦理方面的专著,但是,这些研究并没有系统地讨论 AI 伦理风险防范。国内对于人工智能伦理的研究从近几年才开始起步,大多对于 AI 伦理的研究还不够系统,往往停留在某一方面的研究。部分学者对 AI 带来的伦理问题进行了分析,给出了简单的应对策略。如赵汀阳的《人工智能提出了什么哲学问题》、王绍源的《国外机器人伦理学的兴起及其问题域分析》等。部分学者对某类型的人工智能产品的伦理问题进行探讨,如郭旭芳的《生物医学领域人工智能应用的伦理问题》、王珀的《无人驾驶与算法伦理——一种后果主义的算法设计伦理框架》等。也有学者对 AI 伦理的技术实现进行了探讨,如李伦和孙保学的《给人工智能一颗“良芯(良心)”》等。还有学者从康德哲学角度探讨 AI 行动者的伦理地位,如王东《康德式人工道德行动者的伦理地位研究》。

目前,技术的发展已经远远超过了伦理理论的发展速度,甚至超过了政策的制定速度,呈现出一种技术“倒逼”政府立法、伦理建构的景象。从目前世界各国的相关政策制定上看,政府层面对相关问题的重视程度已然十分高,相关社会问题非常险峻,急需相关政策制度对相关行业的标准、规则、责任等进行明确的规定。学界、技术人员、行业领袖、政府部门需要联合起来,对相关问题进行深入研究 and 探讨,方能为我国 AI 行业发展奠定坚实的基础。如果依旧停留在当下对 AI 部分问题的观点、看法、展望,而缺乏整体性的、全局性的、系统化的 AI 理论和原则框架的研究成果之支撑,这对于我国相关制度的建立是十分不利的。因此,系统化的 AI 伦理理论和原则框架的研究,迫在眉睫。

人工智能的伦理学研究,不仅仅关乎其在实践领域的行为、决策和责任分配等问题,作为哲学研究的前沿性探索,AI 伦理风险防范问题更是关乎道德哲学的基础理论、人工智能的主体性和人之为人的根本等形上问题。仅仅局限于对 AI 的技术、社会问题的表层研究,并不能从根本上解决 AI 带来的社会问题、伦理问题和主体性问题。亟须从 AI 伦理风险防范所涉及的形上理论和形下实践两方面确立研究纲领。理论研究包含 AI 伦理风险防范的道德理论框架和伦理原则体系,从理论上为 AI 伦理规制及风险防范提供道德哲学基础和伦理原则论证。实践问题是从 AI 的主体形态、行为形态和位格形态,分别由道德地位、道德决策和道德责任三大论域,展开 AI 伦理风险防范机制的研究。

第二,亟待建立“理论-实践”“全球性-地方性”相沟通的开放性 AI 伦理研究新思路。

从目前国内外研究现状来看,学界对 AI 伦理问题的研究主要偏向于从“算法”角度对问题进行还原,进而对 AI 的主体本质、伦理风险防范机制进行讨论。如从 AI 底层逻辑中的算法出发,对技术本身进行研究,在此基础上对人机关系进行分析,进而采用技术路径对 AI 的伦理道德进行植入。如安德森夫妇根据生命伦理原则设计的伦理决策系统 MedEthEx (Anderson, M., Anderson, S. L., Armen, C. MedEthEx: A prototype medical ethics advisor),他们以功利主义和义务论为原则开发了两种

伦理决策系统(Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm), 还有 Armstrong S、Dewey D、Wu Y、Abel D 等人通过贝叶斯理论构建相应的效用函数, 以制造能够进行道德决策的机器模型。也有学者从实际问题出发, 从 AI 在实际中的运用着手, 为相应实践中的问题提供一定的理论意见。如白慧仁、隋婷婷等人对无人驾驶汽车事故中的责任问题进行分析, 李德顺、德瑞克·雷本、克里斯蒂安·格德和莎拉·桑顿等对无人驾驶汽车面临两难情况下的道德抉择进行了相关研究, 安德森夫妇、张玉洁、王健、林津如等人对 AI 在医护领域应用中的伦理问题进行了研究。这些研究, 或是单纯地从伦理理论角度对 AI 的本体论问题进行研究, 或是专注于为实践中的伦理规制提供建议, 缺少“理论-实践”之间的沟通性、贯穿性的研究。AI 伦理问题, 既不应该仅仅停留在理论上, 也不应该仅仅从实践问题着手, 这两者都存在着一定的弊端。前者对理论进行深度挖掘, 但在现实应用上往往难以接轨; 后者能够直接对实践问题进行分析, 又往往局限于个别问题、个别事件或个别领域, 难以以为行业整体和长期发展提供建设性的解决方法。

因此, 亟须采用“技术+善法”的问题域还原方法, 从理论上对 AI 的存在论本质进行本体论分析, 构建道德理论框架; 并从 AI 工程学实践的技术类型及其具体分域入手, 结合 AI 伦理嵌入的技术原理和具体情景进行问题探源和机制构建, 以推进“理论-实践”的连通。不仅如此, 由于 AI 伦理的形上理论是本质性的“全球共性”的问题, 而 AI 道德实践不可避免地带有“地方性差异”, 不同国家有着不同文化、国情、伦理风俗, 对 AI 的实践应用也有着深刻的影响。AI 伦理研究, 必须结合我国国情、伦理传统、风俗文化等因素, 推进“全球性-地方性”的沟通; 结合中国历史文化、中国哲学和伦理传统对 AI 道德理论框架进行研究; 探索符合中国国情的 AI 伦理理论原则; 建立“理论-实践”“全球性-地方性”相沟通的开放性 AI 伦理研究框架, 寻求 AI 伦理风险防范的解决之道。

第三, 亟须面向 AI 道德形态探索 AI 伦理研究的新方法。

从“理论-实践”“全球性-地方性”相连通的开放性论题探讨 AI 伦理问题, 需要在研究方法上进行新探索。已有的相关研究中, 学者们通常采用对问题进行问题域还原的方法、学科交叉的方法、实证方法、实验法和思想实验等方法。如伦理学家对无人驾驶汽车的两难困境问题的探讨采用的是思想实验方法。由于目前 AI 应用所产生的案例还很少, 只能通过思想实验或个别的案例进行探讨。也有学者结合实验伦理学方法对相关问题进行研究, 如博纳峰等人采用实验方法研究不同人群、面对不同的情境下对电车难题的选择(The social dilemma of autonomous vehicles), 林芳芳、刘明娟和廖凤林等人从实验角度对电车难题中情感等因素对人们选择的影响进行研究^①。某些学者采用交叉学科方法, 从技术角度实现人工智能的道德植入, 如安德森夫妇的道德决策机器人、瓦拉赫和艾伦对伦理实现路径的探讨等。

上述研究方法能够在一定程度上探寻人工智能的本质、澄清 AI 和人之间的伦理关系, 并给出相应的对策建议。然而, 上述研究方法并不能真正应对 AI 伦理问题的复杂性。尤其是在 AI 伦理规制问题上, “理论与实践”“模拟与现实”“普遍与特殊”“全球性-地方性”往往不可分割地结合在一起, 人们真实面对的是一个不断展开的道德形态过程。如何面向 AI 道德形态过程探索 AI 伦理研究的新方法? 亟须从 AI 道德形态视野切入, 推进 AI 伦理学的方法论创新。

(责任编辑 万 旭)

^① 刘明娟、廖凤林:《电车难题: 情境人数对道德判断与侧隐之心的影响》,《首都师范大学学报(自然科学版)》2013年第2期; 林芳芳:《情感卷入对道德判断的影响》, 硕士学位论文, 广西师范大学研究生院, 2010年。