人工智能治理框架及其人文社会科学研究问题

于昕禾

山东工商学院，山东烟台，264000；

摘要：人工智能(AI) 的不断发展给企业，社会和政府均带来了巨大的机遇与挑战。本文首先通过分析一些知名企业的 AI 治理实践，提炼出 AI 治理“1+7“原则。其次，结合分层分析法和 WSR 模型建立 AI 治理的实施架构。该架构将 AI 治理划分为研发，迭代，运维，问责，训练，共生六大核心活动，技术，泛技术，非技术三大层次，物理，人理，事理三大维度。这次之上，本文研究了由人类，群体，智能体，智能群组，人机混合体，人机混合群组构成的新型复杂社会的人文社会科学特性。最后，针对 AI 治理面临的人文社会科学难点问题，提出了具体的解决措施，以期为 AI 治理提供些思路。

关键词：人工智能；治理；原则；架构；人文社会科学

DOI:10.69979/3029-2735.24.9.050

引言

随着着人工智能技术的突飞猛进，人工智能技术越来越多地被应用于个人生活，商业应用以及社会管理等各个领域。但是要注意的一点是，AI 也带来了很多亟待解决的问题，如隐私泄露，算法歧视，决策失控，失业风险等。这些问题除了对个人、企业有不利影响，也造成了社会普遍的担心。因此，为了更好地把握住 AI 带来的发展机遇，规避潜在风险和挑战。这就需要建立一套科学，全面的 AI 治理体系。文章通过对国内外知名企业的 AI 治理实践进行分析 提炼了 AI 治理原则，构建了 AI 治理架构， 并经过分析新型社会的人文社会科学特点， 为 AI 时代的高质量发展贡献智慧。

1 知名企业的 AI 治理实践与“1+7”原则

1.1 波士顿咨询的实践及启示

作为全球顶尖的战略咨询公司， 波士顿咨询在 AI 治理方面有许多值得借鉴的做法。首先，波士顿咨询在其内部成立了 AI 专项工作组，主要负责 AI 治理框架与最佳实践的制定和推广。其次，波士顿咨询与一些名校，如哈佛大学等展开了深度合作，共同研究 AI 对就业的影响，以此寻找培养人与 AI 协同工作能力的有效路径。此外，波士顿咨询还在 AI 治理领域的前沿观点方面积极发布，如“AI 治理需要平衡创新与规制”，“从道德和价值观层面思考 AI 治理”等观点，引发了业界的广泛关注和讨论。

1.2 微软、谷歌以及脸书等科技巨头的做法和借鉴。

一是微软，其在 AI 治理上采取务实，循序渐进策略。一方面，微软成立“AI 伦理与效果”委员会，负责审查 AI 的重大决策；另一方面，微软在产品开发中严格遵循隐私保护，安全，公平等原则，并积极开展“AI 素养“的培训，借此提高员工对 AI 治理的认知和遵从度。此外，微软还和学术界开展了广泛合作，在 AI 领域资助研究项目。

二是谷歌，其成立了专门的 AI 治理团队，旨在确保 AI 的开发和应用符合谷歌的 AI 原则。同时谷歌还推出了一系列帮助开发者和用户理解 AI 系统的工具，例如“Model Card“可以用来阐明模型的局限性，“What-If“可以用来检测模型可能的偏见。在业务实践中，谷歌也积极顺应 AI 治理的要求，比如允许用户管理自己的数字足迹，提高广告投放的透明度等。

三是脸书，其人工智能治理的重点在于算法的透明度和可解释性。如脸书推出“Why am I seeing this post? 功能，同时，脸书还合作开发了 “Fairness Flow”工具，用于检测模型训练过程中的偏见。另外，脸书还和第三方机构进行了合作， 为一些重大事件识别和删除虚假信息，这体现了科技公司社会责任担当。

1.3 奔驰等传统企业的实践

除了科技巨头，传统企业也纷纷探索 AI 治理。例如， 德国奔驰集团在 2019 年发布了 AI 指南，其共有四项基本原则：负责任的使用， 解释性和透明性，隐私保护，质量和安全。该指南规定，AI 当用于帮助人类决策时，务必要告知用户系统可能存在的局限性。此外，系统决策必须是可解释和可审查的。用户隐私不能被侵犯。AI 的质量和安全必须达到严格的标准。

1.4 AI 治理的“1+7 原则

通过分析以上企业的 AI 治理实践，下面总结出了 AI 治理的“1+7”原则。“1”是指 AI 治理的总目标：让 AI 更好地服务于人类社会的发展。“7”则是指实现

这一目标所需遵循的 7 项基本原则:

① 以人为本:AI 的研发和应用应当以增进人类福祉为出发点和落脚点。

② 公平正义:AI 不应加剧社会的不平等,要公平对待不同的个人和群体。

③ 透明可控:AI 系统的决策过程应当透明,并接受人类的监管。

④ 强化隐私:AI 系统应严格保护个人隐私,未经授权不得收集和使用个人数据。

⑤ 安全可靠:AI 系统必须稳定、可靠,不会给人类生命财产造成损害。

⑥ 包容开放:AI 治理应兼顾不同利益相关方的诉求,鼓励多方参与。

⑦ 问责问责:AI 治理应明确相关主体的责任,对违规行为进行追责。

2 基于分层分析法和 WSR 模型的 AI 治理实施架构

2.1 架构的六大核心活动

我们运用分层分析法和 WSR 模型(即物理(wuli)一事理(shili)一人理(renli)系统方法论)对 AI 治理进行分解,得出了一个三层六棱的架构。在此架构中,AI 治理的核心活动被分为如下几点:

第一,研发:即 AI 系统和工具的研发,包括需求分析、算法选择、模型设计、训练调优、测试验证等环节。

第二,迭代:即对已有的 AI 系统进行改进和优化,使其更加安全、可靠、高效。通常采用增量式迭代开发的模式。

第三,运维:即 AI 系统的日常运行和维护,包括故障诊断、安全防护、版本管理、资源调配等。良好的运维可确保 AI 系统持续、稳定地运行。

第四,问责:即对 AI 系统的研发、使用及其造成的后果进行审查和追责。通常需建立事前审查、过程监管、事后问责的闭环机制。

第五,训练:即对参与到 AI 全生命周期中的所有人(如工程师、管理者、普通员工、用户等)进行相关教育和培训,使其了解并遵守相关政策要求。

第六,共生:即人工智能系统与人类的互利共生,实现“人机协同”。这需要在人机界面、工作流程等方面进行创新设计。

2.2 架构的三大实施层次

在上述六个活动的基础上,架构进一步设置了三个实施层次:

首先是技术层次:也就是 AI 系统本身的技术架构和实现。它直接决定了 AI 系统的性能、安全性、可解释性等。主要涉及研发和迭代两项核心活动。

其次是泛技术层次:即与 AI 系统的开发、部署、应用相关的各项制度流程和保障措施,如伦理审查制度、隐私保护机制、应急预案等。主要涉及迭代、运维、问责三项核心活动。

最后是非技术层次:即更广泛的法律、文化等因素对 AI 治理的影响。比如,不同国家和地区的法律法规存在差异,宗教信仰和价值观念也各不相同。主要涉及训练和共生两项核心活动。

这三个层次相互交织,共同构成了一个立体的 AI 治理蓝图。只有三个层次协同发力,才能真正实现对 AI 全生命周期的有效管控。比如,技术层次的算法透明化设计,需要泛技术层次的审查监管制度予以保障;而非技术层次的伦理道德规范,又会对技术层次的系统开发提出更高要求。

2.3 架构所涉及的三大维度

除了“六棱”、“三层”之外,架构还包含了“三维”。所谓“三维”,是指 AI 治理所涉及的三个基本维度:

1. 物理维度:关注 AI 系统的物质技术基础,包括硬件设施、算力资源、能源消耗等。它是 AI 系统得以运行的物质前提。

2. 人理维度:聚焦 AI 治理对人的影响,关注如何处理就业、隐私、公平等问题。这些问题事关个人切身利益,引发社会的广泛关注。

3. 事理维度:审视 AI 技术发展的内在逻辑,思考其给人类社会带来的深层次影响。比如,AI 可能改变人类固有的认知观念。

在这三个维度中,任何单一的维度都不足以应对 AI 治理的复杂性。所以,必须统筹兼顾,多维聚焦,才能对 AI 的发展形成合力。

3 人、人群、智能体、智能群组、人机混合体及人机混合群组的人文社会科学特性

3.1 人与人群

在 AI 治理下,需要关注人与人群的行为特征和相互关系。一方面,个体的认知、情绪、态度等心理因素会影响其对 AI 的接纳程度。比如,有人会对 AI 抱有不信任感,担心隐私泄露和工作被取代。另一方面,群体内部的互动模式(如从众、极化)以及群体间的对比心理(如偏见、刻板印象),也会左右公众对 AI 议题的看法。这就要求在 AI 治理中充分考虑人性因素,采取有针对性的沟通策略。

3.2 智能体与智能群组

人工智能系统作为一种新型行为主体,有别于传统的人类个体和群体,具有一些独特性质。比如,智能体通

常基于海量数据和复杂算法做出决策,其内在逻辑可能难以解释;多个智能体连接形成的智能群组,其涌现行为更加复杂多变,有时会出现意料之外的结果。这就需要我们发展新的理论视角和研究范式,探究智能体、智能群组及其行为的运作机理和社会影响。

3.3 人机混合体与混合群组

随着人工智能与人类社会的深度融合,诞生了人机混合体和人机混合群组这两种新的存在形态。前者是指人与智能系统的紧密结合,比如脑机接口就使人脑与计算机实现了直接联通。后者则是由人类个体、群体、智能体共同构成的一个复合系统。这些新的存在形态突破了人与机器的二元对立,呈现出独特的人文社会科学特征。比如,人机混合群组的决策机制、组织形态、文化心理等,可能与传统的人类社会有很大不同。对此,人文社会科学研究需要与自然科学、工程技术学科深度合作,开展跨学科的探索。

4 AI 治理所面临的人文社会科学问题及解决方案

4.1 主要问题

基于对新型社会形态的分析,AI 治理在人文社会科学领域面临如下问题:

一是个体层面:如何保障个人隐私和知情权?如何提升个体对AI的信任度?如何赋予个人更多的参与度和控制权?

二是群体层面:如何处理不同群体在获取、使用 AI 方面的数字鸿沟?如何防止算法偏见加剧既有的社会不平等?如何在群体中形成理性、包容的 AI 治理共识?

三是人机关系:如何厘清人与 AI 的权责边界?如何实现人机协同,优势互补?如何预防和化解人机冲突?

四是社会层面:AI 的广泛应用将带来哪些深层次的社会变革?如何评估和引导这些变革?如何重塑教育、就业等社会制度以适应智能时代?

五是伦理道德:如何界定 AI 系统的道德地位?如何避免 AI 被用于非法或不道德目的?如何处理 AI 可能带来的伦理难题,如隐私、歧视、操纵等?

4.2 解决方案

其一,加强跨学科研究。AI 治理所涉及的问题十分复杂,单一学科难以应对。故必须打破学科壁垒,加强计算机、法学、社会学、心理学、伦理学等跨学科交叉研究,形成合力。如可成立“AI 治理研究中心”,聚集不同背景的专家学者开展协同攻关。

其二,完善法律法规。要加快 AI 治理领域的立法进程,尽快出台专门的法律法规,明确政府、企业、个人等各方责权利。同时,要与时俱进地修订现有法律,使之适应 AI 技术发展的需要。比如,要重点加强个人信息保护、算法歧视治理等方面的立法。

其三,强化伦理教育。伦理道德是 AI 治理的重要底线。要将 AI 伦理教育纳入到计算机、数据科学等相关专业的培养方案中,提高从业者的职业道德修养。在全社会广泛开展 AI 科普教育,提升公众对 AI 的理性认知,形成积极健康的 AI 文化。

其四,发展监管科技。应运用先进技术手段,加强对 AI 系统的全生命周期监管。如利用区块链技术,对数据采集、算法开发、系统应用全链条进行追溯管理;开发AI审计工具,对算法模型进行偏见检测和去偏处理;搭建 AI 沙盒环境,对新系统进行封闭测试,评估其风险和影响。

其五,促进多方参与。AI 治理是需要政府、企业、学术界、公众等多元主体共同参与的。所以,需搭建开放包容的合作平台,鼓励不同利益相关方加强对话交流,在博弈中凝聚共识。此外,要营造宽松的制度环境,调动全社会参与 AI 治理的积极性和创造性。

结语

综上所述,本文以知名企业的实践为基础,总结提炼了“1+7”AI 治理原则,并构建了一个多维立体的 AI 治理实施架构,涵盖了六大核心活动、三大层次、三大维度。接着分析了人机共生时代的新型社会形态及其人文社会科学特点,针对 AI 治理面临的主要问题,提出了一些解决方案。当然,AI 治理不是一蹴而就的,需要在实践中不断探索完善。今后,还应在更大范围、更深层次上推进 AI 治理研究,为人工智能健康发展保驾护航。

参考文献

[1]刘露,杨晓雷,高文.面向技术发展的人工智能弹性治理框架研究[J].科学与社会,2021,11(2):15-29.
[2]朝乐门.人工智能治理框架及其人文社会科学研究问题分析[J].情报资料工作,2022,43(5):6-15.
[3]周江伟,赵瑜.人工智能治理原则的实践导向:可靠性、问责制与社会协同[J].治理研究,2023,39(5):111-127159,160.
[4]王阁.人工智能的全球治理背景及其治理框架[J].实事求是,2024(5):92-100
[5]杨永恒.人工智能时代社会科学研究的“变”与“不变”[J].学术前沿,2024(4):96-105.

作者简介:于昕禾,2004.11,女,汉族,本科,研究方向:工商管理。