人工智能技术的潜在社会风险及治理体制机制研究

丁元竹

[摘　要] 人工智能安全及其风险治理，包括人工智能的社会治理等问题，已经成为国内外各界共同关注的话题。引发人工智能风险议题的主要原因包括：人工智能技术仍处在探索和发展阶段、技术激进主义造成认知误区、科学技术与人文社会科学的互补机制不健全、“技术能力”与“社会期待”之间“错位”、监管体系不够完备等。构建可控的人工智能治理体制机制，可以从完善人工智能发展的评估内涵工作、完善跨学科综合研究开发体制机制、建立沙盒监管机制、建立人工智能负责人制度、建立人工智能治理的公众参与机制、加强人工智能领域国际合作的体制机制等方面着力。

[关键词] 人工智能；潜在风险；治理体制机制
[中图分类号] D63　　[文献标识码] A　　[文章编号] 1674-7453(2025)07-0004-13

引言

本文所讨论的人工智能，主要指生成式人工智能（Generative Artificial Intelligence），以及未来将会出现的通用人工智能（Artificial General Intelligence）和超级人工智能（Artificial Super Intelligence）。当前，人工智能正处在由实验室研发向产业化运用快速转型阶段，何时成熟取决于其迭代速度。

人工智能的治理将会是一种新型且十分复杂的治理形态。人工智能安全及其风险治理，包括人工智能的社会治理等问题，已经成为国内外各界共同关注的话题，有关它的讨论越来越多，也越来越激烈。这些讨论包含了如何认识人工智能技术的本质及其对人类命运的影响等一系列深层次问题，尤其是开发方式、监管措施，以及与之相适应的文明形态等。其实，人工智能的真正挑战不在于技术本身，而在于如何把它接入人类社会和文化的体制机制。未来，发生在人工智能领域的竞争不仅是算力竞争，更是制度体制和文化适应性的较量——就像电动车的成功既依靠电池技术，也依靠充电网络一样。“人工智能之所以特别，以至于有必要调整私法来促进其推广并挖掘其增

[作者简介] 丁元竹，中共中央党校（国家行政学院）教授、博士生导师。

长潜力，原因在于它具有许多新的特性，其中两个尤为重要：自主性和不透明性。”[1]在这个意义上，讨论人工智能的治理就需要摆脱既往的思考习惯，必须基于治理与技术同步的相关体制机制研究，探索基于“自主性和不透明性”治理的底层逻辑：AI不是简单的技术工具，而是在受人类大脑启发基础上开发的具有智能特征的事物，管理好人工智能是社会治理的重要内容和基础性工作。未来，大模型实体化、垂直化，具身智能社会化将会涌现，坚持以人为核心的人工智能发展，使人类成为AI培训师、引导者，实现人机良性互动将成为人工智能治理的中心问题。因此，人工智能的治理是技术与制度、体制机制、社会文化的发展进化和适应过程。

一、关注进入赋能阶段的人工智能技术及其风险议题

人工智能技术正在以前所未有的速度发展，广泛渗透到各个领域，重塑人类生产方式、生活方式，以及社会运行等。但是，在这场看似璀璨的科技变革浪潮之下，却潜藏着一系列不能忽视的暗流。尽管人工智能技术尚未达到高度自主可控水平，但随着技术不断发展迭代，未来不排除出现技术失控的可能性。因此，有必要从现在开始高度重视人工智能技术发展过程中可能带来的风险与挑战。

（一）人工智能技术已进入赋能阶段

当前，尽管大家在日常生活中还未全面感觉到人工智能已经彻底改变了我们的生活，但人工智能大模型从实验室技术进入赋能和落地应用阶段已经成为事实。一是大语言模型已经可以开展多轮对话、形成代码、进行文本创作等运行。百度已发布端到端语音语言大模型，实现了超逼真语音交互。有关情感的技术开发也在进行中，MiniMax Audio推出新模型Speech-02，支持17种语言及300多种真实音色，可一键将长达20万字的文档转为语音；平台提供情感设置选项，包括高兴、生气、伤心等8种情绪，还可调整声音深度和强度，音色自然度高且情感丰富；用户仅需10秒录音即可1:1克隆声音，每日赠送4000积分（约5分钟音频），适用于跨境电商、AI出海、角色扮演等场景①。二是在专业领域的应用取得明显进展，如，Harvey（服务于律所的专用人工智能）可辅助完成合同分析等事项。在医疗领域，人工智能已经能够实现协助诊断肺癌、糖尿病视网膜病变等疾病。达特茅斯团队研发的AI心理治疗机器人“Therabot”临床试验显示可使抑郁症状平均减轻51%，疗效堪比人类治疗师②。AlphaFold服务器（DeepMind）可以预测超2亿种蛋白质结构，加速了新药研发进程。人工智能个性化医疗技术可以提高癌症治疗效果，瑞典卡罗林斯卡学院研究团队开发的DNA纳米机器人在小鼠实验中实现了成功抑制70%肿瘤生长，通过精确激活仅在肿瘤微环境中的“武器”实现靶向治疗；纳米机器人采用脱氧核糖核酸（DNA）折纸技术构建，能在癌细胞特有的低pH值环境中触发，展开结构释放细胞毒性配体，诱导癌细胞凋亡；研究展示了脱氧核糖核酸纳米机器人在癌症治疗中的潜力，为未来人类临床试验和癌症治疗方法提供了新方向③。人工智能工业机器人分拣系统使工作效率提升了数倍。人工智能在金融科技（FinTech）领域协助金融系统优化投资组合，及时发现金融欺诈和抑制欺诈等交易。三是在多模态交互中人工智能支持图像理解，实现文本+图像+音频联合推理等目标。用于受控数据集实验的测试平台（DCLM）团队清

①MiniMax Audio最新语音模型，一次性可以输入20万字符。来自腾讯研究院AI速递，20250403。

②抑郁症状平均减轻51%，首个AI心理治疗师临床试验报告出炉。来自腾讯研究院AI速递，20250401。

③人机融合即将成真！纳米机器人杀死癌细胞，肿瘤生长抑制70%。详见https://news.qq.com/rain/a/20240703A04KB000。

洗出240万亿巨量数据，足够训出18个GPT-4，强调不再单纯增大模型规模（Scale Up），而是通过数据质量提升和优化（Scale Down），推动下一代AI模型发展[①]。特斯拉使用人工智能技术检测车身存在的问题，准确率高达99%以上。不少国家和地区的无人驾驶出租车在人工监督情况下实现城市运行。杭州亚运会期间，“智慧高速”通过人工智能优化交通流量，提高了城市出行效率。

简言之，人工智能已经进入“深度赋能”的快车道，从提升生产效率、创新商业模式到重塑科学研究模式，人工智能的应用无处不在。2023年，全球活跃互联网用户超过53亿，占全球人口的65.4%，这意味着全球超过65%的人口可以上网，预计2025年互联网用户数量可能达到65.4亿[②]。这为人工智能布局奠定了基础。未来，随着多模态大模型开发、量子计算+人工智能创新、脑机接口技术发展、具身机器人布局市场，人工智能的影响力必将进一步扩大。哈佛团队在Nature发文综述生成式医学图像解释（GenMI）领域进展，指出AI系统有望通过视觉编码器和语言解码器自动生成跨学科医学报告；生成式医学图像解释可通过“AI住院实习医师”和“符合人类偏好”两种范式与临床医生协作，帮助诊断、快速检索类似病例和辅助决策；实现生成式医学图像解释潜力仍需克服四大挑战：建立可靠评估基准、解决临床医生过度依赖、消除数据集和模型偏差、扩展到三维影像和其他医学专科领域[③]。人工智能技术突飞猛进，但统筹创新与安全之间的关系仍然且必将是人类长期面临的重大课题。人工智能发展已不可逆，它的终极前景取决于人类如何引导它的发展，与之有关的人类价值观和制度设计将决定着未来的人工智能成为“普罗米修斯之火”还是“潘多拉魔盒”。

（二）人工智能赋能带来的风险议题

随着人工智能技术的不断进步，统筹其发展与安全被日益前置化、动态化和精细化，人们在技术开发初期就开始考虑安全因素，根据不同的场景和风险等级设置差异化治理策略。在这样的背景下，产生了一系列风险议题，下面列举部分主要议题。

1.人工智能是否会超越甚至替代人类。对于人工智能的未来，人们态度各异。有人认为，如果人工智能无法与人类价值观、伦理规则和社会大多数人的利益保持一致，会引发个体权利受到侵害，甚至对人类文明形成严重威胁。前Open AI研究员丹尼尔·科科塔伊洛（Daniel Kokotajlo）团队发布了76页“AI 2027”报告，预测通用人工智能将在2027年中期实现；报告描绘出从2025年最贵AI诞生到2027年Agent-5渗透政府决策的发展路径，预计AI将快速超越人类智能；研究团队预测超人AI的影响将超过工业革命，不过也有部分专家质疑这种预测缺乏科学依据且过于极端[④]。有关人工智能未来的风险可以归结为可能出现目标错配问题。人工智能根据设定目标严格预设指标，但它无法理解人类价值观的复杂性和模糊性。人工智能军事系统或许会擅自升级冲突，如无人机作出误判，视无辜的平民为威胁目标并发动攻击。超级人工智能企业通过算法垄断定价权、操纵市场，损害消费者利益。美国COMPAS司法系统对黑人被告给出“再犯罪风险更高”的误判，强化了种族之间的偏见，甚至引发种族之间的冲突。算法的个性化推送制造了“信息茧房”，造成社会

①240万亿巨量数据被洗出，足够训出18个GPT-4！全球23所机构联手，清洗秘籍公开。详见https://www.36kr.com/p/2833371245775107。

②2024年全球互联网用户统计数据。详见https://zhuanlan.zhihu.com/p/682732049。

③Nature：AI击败人类医学专家？哈佛团队：这一领域仍需解决4大难题。详见https://www.huxiu.com/article/4167669.html（哈佛论文链接：https://www.nature.com/articles/s41586-024-07618-3）。

④“AI 2027”预测报告，2027年实现自我进化的Agent-5。来自腾讯研究院AI速递，20250408。

舆论的极端化和社会矛盾加剧。过度和长期依赖人工智能会导致人类判断力、思考力退化，如飞行员会因长期使用自动驾驶技术造成驾驶技能生疏，在遇到需要处理的紧急情况时措手不及。人工智能价值对齐（AI Alignment）是个问题。人工智能价值对齐是确保人工智能系统的目标、决策和行为与人类价值观、伦理规范以及社会利益保持一致的核心原则。人工智能价值对齐为什么如此困难？是因为人类价值观的多元性与矛盾：不同文化对“公平”“自由”有着自己的理解和定义。人工智能价值对齐不仅是对技术的挑战，更是对文明存续的考验。人工智能造福人类的关键在于把价值对齐视为一个持续过程，而不是一次性解决方案，如尽管通用人工智能的出现不可避免，但通用人工智能发展不仅取决于大模型技术，也取决于机械技术进化。

2. 人工智能是否会严重冲击就业市场。人工智能对就业市场的冲击也是人们最为关注的议题之一，认为其替代人类工作的速度和范围或许会远远超过以往任何一次技术革命。目前看，具有高替代风险的岗位是那些规则明确型的工作，诸如制造业（像装配线工人）、会计（如自动化报表人员）、基础法律文书（包括合同审核人员），等等。富士康已经用机器人替代了超过60万人的重复性工作岗位。数据处理型工作，包括客服（像聊天机器人）、初级医疗诊断（如医疗影像识别人员）、金融分析（包括算法交易人员）等。在图像处理领域，字节跳动发布DreamActor-M1视频生成框架，只需一张参考图像就能模仿视频中人物行为，实现高质量人体动画[①]。高盛纽约现金股票交易员已经从2000年的600人减少至2023年的2人。具有中等替代风险的岗位包括，部分创意工作（像广告文案人员）、平面设计、新闻写作、音乐作曲，以及个性化辅导等。那些情感互动性高或复杂性程度高的决策工作，诸如心理咨询师、高级管理人员、社会工作者、园艺师、养老护理人员等目前被替代的可能性和数量或许不会太大。比尔·盖茨（Bill Gates）认为，医生和教师等职业将被AI取代，但软件开发者（主要指发现和修复错误、优化算法等）、生物学家和能源专家有望幸存[②]。而且，不容乐观的是，行业转型并非易事，那些被替代的流水线工人或文员很难快速转型为人工智能训练师或机器人维护员，2022年美国仅54%的失业者成功转入新行业。

3. 人工智能是否会造成更严重的收入两极分化。眼下，高技能人工智能岗位，如算法工程师薪资暴涨，而低技能岗位则不断萎缩，中等收入群体的工作岗位出现空心化。从个人计算机和互联网广泛普及以来的历史看，在缺乏相应制度机制介入的情况下，经济不平等会进一步扩大，掌握人工智能技术的资本所有者与普通劳动者之间的收入差距会持续扩大。大量失业可能导致自我价值感丧失、精神焦虑和抑郁症人口的比例上升，社会问题会凸显。由于目前无法问责算法、特定人群受到不公平待遇、一个人意外死亡，这些都是人工智能领域可能出现的新局面。[2]未来，算法差异或许会成为社会文化鸿沟和不平等的根源。

4. 人工智能应用中存在的数据安全风险。大模型有可能保存、记忆，并泄露训练数据中的敏感信息，如个人病历、企业信息等，它也可能通过健康咨询记录推断用户疾病类型和状况，并推送某些针对性治疗广告。对抗攻击（Adversarial Attacks）通过输入篡改，如轻微修改图像或文本会造成自动驾驶出现误判，进而造成交通事故。攻击者通过应用程序编程接口反复查询人工智能系统，复制其数据信息，如复制人脸识别系统，可能带来社会风险。部分不法分子恶意注入错误数据，如通过垃圾邮件分类器投喂正常邮件并标记

①DreamActor-M1替代动捕？基于DiT的人体动画生成框架。来自腾讯研究院AI速递，20250407。

②盖茨预警：医生和教师最先被取代，未来一周只上2天班。来自腾讯研究院AI速递，20250401。

为“垃圾”，可能造成大模型失效。也有人通过大模型输出推测某些敏感属性，如通过贷款审批人AI去推断用户种族特征等，可能造成隐私泄露问题。地理空间数据泄露可能会暴露军事设施位置，造成军事攻击行为。未来，量子计算技术有可能破解现有加密方法，威胁人工智能数据安全。上述现象表明，人工智能数据安全不是单纯的技术问题，而是涉及法律、伦理、经济和社会的系统性安全问题。数据是人工智能时代的石油，但当它发生泄漏时则更像放射性物质，会给人类带来巨大灾难。

5.神经网络目前的不可解释性难题。深层神经网络基于数百万甚至数十亿参数的非线性组合作出推理和决策，就目前的技术能力而言，人类很难追踪它的逻辑路径。卷积神经网络的深层神经元或许能够识别“纹理”“形状”等抽象特征，但这些特征与人类认知架构并不能直接对应，具有不可解释性。现有大模型从数据中学习，不依赖人工预设规则，也不依赖人类的社会互动方式，无法解释为什么它会有“这样的”答案，而不是“那样的”答案。人类可以根据已有的理论、历史、经验和事实进行决策判断和未来预测。而目前的人工智能大模型很难做到这点。欧洲通用数据保护条例（General Data Protection Regulation）规定用户有权获得“自动化决策的解释”，但现有人工智能系统无法满足这样的要求。美国食品药品监督管理局（U.S. Food and Drug Administration）要求医疗人工智能必须提供“可验证的决策依据”，事实上很多深度学习模型无法通过这类审批。事后解释工具，如使用训练的局部代理模型来对单个样本进行解释（Local Interpretable Model-Agnostic Explanations）和比较全能的模型可解释性的方法（SHapley Additive exPlanation）只能提供局部的、近似的解释，还无法还原真实决策的逻辑。注意力机制，如转换架构（Transformer）可以显示大模型“关注”的输入区域，但无法解释“为何关注”。简单模型，如决策树具有解释性，但性能不高；复杂大模型性能高，但具有很大的不可解释性。目前，神经网络的不可解释性既是人工智能技术发展的瓶颈，也是处理好人机关系的关键。未来，在人工智能领域，人类需要在“有限可解释性”基础上通过制度与技术协同提高风险管控水平。

6.大模型训练可能造成资源过度消耗和生态环境破坏。大模型训练依赖庞大的计算资源，需付出高昂的经济、环境成本，对全球资源开发、保护、分配、能源结构和生态环境产生重大影响。人工智能有可能在未来十年内改变能源行业格局，推动全球数据中心电力需求激增。[3]一般来说，训练大模型需要数千张甚至数万张，乃至数十万张的高端GPU/TPU，如英伟达的H100、谷歌的TPU v4等，这导致全球算力资源向少数科技巨头集中和财富向个别基础设施生产商集中。先进芯片，如3nm及以下制程依赖台积电、三星等少数厂商，在地缘政治上加剧了算力垄断。巨量算力投入使学术机构和小公司难以参与研发竞争。单次训练的高能耗加剧了区域的能源压力。芯片制造与冷却的巨大用水量带来巨大的环境资源压力。大模型高度依赖互联网公开文本，由此可以预见，优质数据（包括书籍、学术论文等）极可能在未来会耗尽，而合成数据可能导致大模型“近亲繁殖”，性能下降。未经许可的社交媒体、版权内容抓取带来了一系列版权和伦理问题。人工智能芯片，如英伟达每两年升级架构，淘汰了大量过时的GPU，形成海量电子垃圾。先进制程芯片依赖钴、锂、稀土等自然资源，对这类资源的开采将加剧生态环境的破坏，造成发展的不可持续性。

二、引发人工智能风险议题的主要原因

人工智能迅猛发展，其潜在风险已成为世界各国学术界、产业界和政策部门关注和防范的焦

点之一。“技术的发展扩大了‘可能性空间’,例如飞机的发明使载人飞行成为现实,而可能性空间的扩展则带来了正反两方面的影响。”[4]人工智能风险非单一因素所致,而是由其技术的特性、市场契合度、社会适应性、相关制度体制及社会认知、文化习惯等诸多因素交织而成。错综复杂的因素交织造成了人工智能的潜在风险,系统全面剖析这些原因,有助于采取针对性措施,确保其健康、有序、可持续发展。

(一)人工智能技术仍处在探索和发展阶段

总体来看,人工智能技术还不太成熟。尽管人工智能在过去十年中取得了不可思议的进展,但很多研究仍处于起步阶段,尤其是机器人技术,这项技术是出了名的难,即使在深度学习普及的时代,机器人技术的发展步伐也相对缓慢。[5]一是从技术发展阶段性看,当前的人工智能在抽象推理、常识理解、情感表达、自我意识、创造性思维等方面还远未达到人类的水平。例如,大语言模型会生成文字流畅但逻辑错误的答案,无法真正理解上下文和真实的语义。2025年4月初,Llama 4发布36小时后遭遇大量负面评价,尤其在代码能力测试中表现不佳,经典“小球反弹”测试失败①。大模型的准确性依赖高质量数据,在面对数据资源稀缺、数据偏差或动态变化的环境时,大模型表现出极大的不稳定性,这种情况也出现在医疗诊断、复杂决策等过程中。二是从应用场景看,部分领域的人工智能应用已高度成熟。人工智能在特定任务上已经超越人类,如图像识别,包括医疗影像分析、语音识别、围棋对决、推荐系统(包括电商、短视频)等。许多企业通过人工智能优化业务流程,这类“窄人工智能”技术已成熟到可以规模化和落地程度。有些大模型虽然不尽完美,但可以成为生产性工具,进入生产过程和市场领域。三是期望与现实之间存在差距。公众对人工智能的认知时常被媒体夸大,造成一定程度的技术迷茫。技术发展快于法律和伦理框架的建设也增加了人工智能的“不成熟”。“技术表现接近‘人类水平’的说法本身,会让人觉得是臆想,甚至是海市蜃楼。人类的能力维度是丰富多样的,远非任何单一指标所能衡量。但我们的缺点和优点一样具有启发性。”[6]只有更好地了解和认识人类自身,才能更好地认识人工智能。“人的感知能力虽然有种种局限,但与机器截然相反。我们从整体上看待世界,不仅能识别世界的内容,更可以进一步理解不同事物之间的关系、意义、过去和未来。”[7]迄今为止,人类对自身的认识仍然存在诸多空白。

(二)技术激进主义造成认知误区

技术激进主义(Technological Radicalism)是一种主张快速、大规模应用新技术的社会思潮,它的核心思想是,技术可以解决社会、经济和环境等多方面存在的问题。技术激进主义的核心观点包括:迅速推广新技术,技术进步是解决社会问题的关键,技术能够超越传统社会、伦理和环境限制。有些技术激进主义者甚至把技术视为社会发展的决定性力量。技术激进主义往往忽视了技术带来的社会、经济和环境风险,单纯追求技术的短期效益。现代意义上,技术激进主义的兴起与20世纪中后期的科学技术革命密不可分,尤其在信息技术、生物技术和人工智能等领域取得突破性进展后,人们对技术的期望值显著提高,认为技术可以重塑社会和经济结构。随着经济社会问题复杂化,人们寄希望于技术提供快速解决各种问题的方案,如通过人工智能解决就业问题,通过基因编辑解决疾病问题,通过航天技术解决生态灾难带来的人类生存和延续问题,等等。

(三)科学技术与人文社会科学的互补机制不健全

当前,由于相当一部分人文社会科学研究者缺乏对现代科学技术的深刻理解,同时相当一部分技术研发人员又缺乏人文社会科学的深厚素

①Llama 4发布36小时差评如潮!匿名员工爆料并拒绝署名。来自腾讯研究院AI速递,20250408。

养，使人们对人工智能的发展方向问题产生了一定程度的争论，甚至出现了认知偏差。缺乏丰富技术涵养的人文社科研究者难以提出切实可行的伦理和法律框架，导致人工智能发展中出现方向性偏差。技术研发人员忽视人工智能的人文社会因素，导致算法偏见、隐私侵犯等问题。技术人员忽视用户体验、市场契合度、社会接受度，导致技术难以更好地融入市场和社会。算法推荐有时强化偏见，影响社会价值观，威胁公共安全。大模型生成的内容在不适合的文化环境中传播，容易引发冲突。技术人员如果忽视人工智能对就业市场冲击的考量，就会缺乏应对策略。而人文社科学者缺乏对技术规律的认识，就难以提出有效的经济社会政策，导致技术与社会脱节。

（四）“技术能力”与“社会期待”之间“错位”

这里需要提出媒体传播“恐怖谷效应”，它指的是媒体对人工智能威胁论（如造成大规模失业）和奇点论的过度渲染，形成“要么乌托邦，要么反乌托邦”的极端叙事。实际上，技术演进更多呈现渐进式、改良性特征。资本驱动“期望膨胀曲线”使一些初创公司为融资目的刻意营造技术突破的假象，造成社会预期过高。人类倾向于把对话流畅性等同于思维深度，导致对聊天机器人产生不合理的心理预期，带来社会焦虑。

自2022年初人工智能大模型爆发引发全球关注以来，人工智能确实面临着“技术能力”与“社会期待”之间的错位。这种认知偏差既源于技术突破带来的极度震撼，也暴露了公众对人工智能发展规律和发展实际的普遍误解。树立“技术成熟度曲线”意识，以区分实验室技术与商业产品可用性之间的差异非常关键，通常，前者会把实际效用夸大到超出人们的想象。要发展“批判性技术素养”，提高人们理解概率模型本质的能力，即当大模型说“明天有95%的可能性下雨”时，并不意味着它掌握了天气变化的规律，而只是基于历史上的统计数据进行的推断，而不是基于理论、数据、经验、案例、历史等诸多因素的综合判断。当前，人工智能正处于“期望膨胀期”向“理性沉淀期”过渡的关键阶段。技术的真正社会价值会在泡沫消退后才能显现出来，人工智能技术也不例外。真正的学者要揭示技术发展史上反复出现的“创新幻觉周期”，帮助人们清醒地认识眼下技术进步的真实情况，这是避免集体非理性的关键。当前的挑战是，要在保持发展技术热情的同时，培育出能准确丈量人工智能真实能力的“社会标尺”。首先，要摆脱“流畅性陷阱”的迷惑性。的确，现行大语言模型生成文本的速度之快、语言之流畅、层次之分明令人感叹，但它缺乏真正的因果推理能力，尤其是缺乏人类具有的内省和意识能力。例如，大模型可以撰写出哲学论文框架，却无法真正理解“自我意识”的本体论意义，这种“语义cosplay”常被误认为智能涌现。人工智能在特定封闭场景，如围棋、蛋白质折叠等领域表现出卓越的能力，但面对开放世界的复杂变量时仍显得力不从心，如自动驾驶汽车在暴风雨中的决策失误率比人类高出数倍。尽管采用人类反馈强化学习（Reinforcement Learning from Human Feedback），大模型的价值取向依然受到训练数据所隐含的偏见影响。

（五）监管体系不够完备

人工智能系统缺乏一整套规范严格的监管制度，究其原因，技术发展速度远超政策制定、伦理共识的发展和完善，以及治理框架本身的复杂性。一是技术迅速迭代，立法相对滞后。当前，人工智能技术更新周期以月或季度计算，甚至以周计算，而法律和监管制定往往需要数年时间。2022年初，ChatGPT的爆发式应用使各国监管机构措手不及。人工智能涉及算法、数据、算力等诸多维度，监管部门缺乏足够的知识储备和技术专家参与政策讨论，致使规则难以精准聚焦风险。二是大模型应用场景广泛，属于通用技术，难以制定统一标准。医疗人工智能、金融风险控

制、军事人工智能应用等领域的风险等级和伦理要求截然不同，一刀切式的监管会扼杀创新或造成漏洞。人工智能产业链全球化，美国的芯片、中国的数据、欧洲的用户等，造成各国监管制度之间的冲突。三是核心问题上的争议难以达成共识。如果自动驾驶汽车出现交通事故，责任属于开发者、运营商还是用户？这些问题都难以套用现有的法律法规去解决。在医疗领域是否允许牺牲少数患者数据隐私提升多数人的治疗效果？这涉及各个文化对隐私的理解和效用的权衡。大模型生成内容的版权归属、人工智能是否应享有“电子人格”等争议在目前依然没有定论。四是利益相关方之间的激烈博弈。科技巨头主导话语权，大型人工智能企业通过游说影响政策，倾向于“自律”而非外部监管。各国，尤其是一些大国，担忧严格监管会削弱技术竞争力，导致以政策妥协牺牲政策效果或延迟政策实施，等等。五是现有监管政策存在局限性。多数政策是在问题爆发后才出台，缺乏政策和法规的前瞻性。要求人工智能系统“完全透明”目前在技术上还不可行，因为深度学习存在黑箱问题，或导致企业负担过重，或影响发展创新。

三、构建可控的人工智能治理体制机制

就目前人工智能的应用状况看，人工智能社会治理更准确地说是一个属于未来的话题，但当下的关注和探讨将直接决定着人类能否在人工智能技术爆发时保持主动。因此，从技术、法律、伦理、教育、社会协作、国际协同等多维度提前布局十分必要。最终，能否实现“人工智能向善”，取决于人类当下的选择。站在人类文明演化的十字路口，建立包含技术在内的各种防火墙已刻不容缓，我们有必要在算力爆炸与人性坚守的张力之间，构建内含包容、反思、韧性的智能社会。

（一）完善人工智能发展的评估内涵工作

我们首先要探讨的问题是，人类需要的是超越自己的人工智能，还是协助人类提高自己福祉的人工智能？成熟的人工智能既是技术飞跃，更是人类社会的认知延伸和合作助手，它具备的核心特征将涵盖技术、伦理和社会。一是技术层面。通用人工智能可以像人类一样跨领域学习，从医学诊断到艺术创作，它无需针对每项具体任务重新训练，且具备因果推理能力，能够理解“为什么”而不仅是“相关性”。成熟的人工智能具有自主与协作平衡功能，可以在独立完成复杂任务的同时实现与人类行动无缝协作。它还能突破现有算力限制，实现类似人脑的低功耗、高效能。在这方面已经取得了一些进展，加州初创公司 Lightmatter 推出光子超级芯片 M1000，提供 114 Tbps 总光带宽，能在单一域支持数千 GPU 互联；公司发布全球首款 3D 共封装光学产品 L200，通过无边缘 I/O 技术实现整个芯片区域带宽扩展，性能提升 5 至 10 倍；创始人尼古拉斯·哈里斯（Nicholas Harris）利用光子计算重塑芯片通信方式，克服电 I/O 连接仅限于芯片边缘的带宽限制，希望解决 AI 发展瓶颈[①]。二是认知与交互类人化。如果人工智能具备社会常识和情感共情，就能够识别用户情绪并调整沟通方式，处理隐喻、幽默，理解文化之间的差异，创造性地解决问题，提出人类想不到的解决方案。三是它的伦理与安全具备完备性、可解释性与透明性，决策过程像“教科书”一样透明清晰，能动态适应不同文化和伦理要求。四是具备深度社会融合能力，它赋能人类而非替代人类，专注于人类不擅长或那些重复性的工作，协助增强人类的创造力。它具备经济与法律身份，拥有承担有限法律责任的能力，如人工智能医生需要通过职业考试获得执业资格，作出错误诊断需要赔偿。通过动态进化，人工智能持续学习和更新自己的

① 实现最快 AI 互连？加州初创光子超级芯片支持数千 GPU 互联。来自腾讯研究院 AI 速递，20250408。

知识库，但不能篡改其核心伦理架构。五是潜在的文化影响。人工智能与人类合作共同创作突破传统媒体的作品，如沉浸式人工智能戏剧，实时改变观众情绪的剧情等。这里就产生了人工智能悖论——“完美的不完美”，即使人工智能技术成熟，它可能仍会存在“缺陷”，如偶尔出现创造性错误，但会维护人类的主导性和多样性价值。成熟人工智能的终极标志应当是，人们将不再讨论人工智能是否成熟，而是像用电一样随意使用人工智能。当然，“最迫切需要AI快速发展的领域都是攸关人命的领域，这是人类发展的必然过程。所以我们必须努力让人们不再排斥AI。”[8]随着人工智能技术快速发展，价值对齐引起了人们的广泛关注。通过价值对齐，可以增强用户信任，开发应用场景，推动技术落地。价值对齐需要解决在多元文化和多元价值观中寻求共同价值的共识，避免算法歧视或偏见。解决好这个问题，需要所有利益相关方的参与协作，需要科学家、技术专家、社会学者、人类学者、伦理学者、政策制定者及公众共同制定对齐标准，还要建立动态反馈机制，通过用户反馈和问责审计不断校准人工智能的行为。人工智能价值对齐是人类社会的“镜像”——它要求人们在智能时代重新审视自身的价值观体系。唯有实现技术迭代、伦理深思和公众参与的结合，才能构建既强大又可控的人工智能。从长远看，解决人工智能的可靠性、可解释性、数据偏见等问题需突破认知科学、生命科学、能源效率（如降低算力依赖）、人机协作等领域的问题。如果以“通用人工智能”或“超级人工智能”为标准，目前的人工智能技术确实不成熟。但如果认为“所有人工智能都不成熟”，则会忽略特定领域的人工智能已经产生了落地效应。例如，北京大学等研究团队通过在适当大小的数据子集上多轮预训练，提出改善大语言模型在医疗领域的训练效率和性能；介绍了使用高质量医疗数据进行连续预训练，并通过混合数据训练缓解预训练分布差异；通过三种策略优化的Llama-3-Physician-8B小模型在医疗任务上表现优越，接近GPT-4级性能①。

（二）完善跨学科综合研究开发体制机制

各类大模型在探索解决价值对齐问题时都需要考虑不同学科之间的合作，尤其是要考虑人工智能技术与人文社会科学之间的合作。人工智能时代，文明正在经历着迭代发展，认知鸿沟的本质是工具理性与价值理性在智能时代的失衡。当人工智能介入意义生产、重塑社会关系，如调解家庭矛盾时，单纯强调技术主义或人文主义都将导致文明的畸形发展。在这个意义上，人类需要构建新的“认知界面”，技术人员需要有诗性思维，算法不仅是数字表达，更是承载着社会价值和文化基因的意义载体。人文学者需要意识到伦理原则必须转化为可执行的验证协议。人工智能时代，唯有打破学科藩篱，实现跨学科合作，人类才能实现真正的“诗意的栖居”。这不仅是知识体系的融合过程，更是人类认知范式新的启蒙运动。探讨用一种文化价值作为算法公平性评估维度，确实需要谨慎。多元文化内涵和现实复杂性是人工智能发展不可回避的课题。用单一文化尝试大模型设计既有其文化合理性，也面临着过于简化带来的风险，成熟的大模型需要从多维进行辩证分析。“人工智能的未来仍然充满不确定性，我们有很多理由保持乐观，也同样有很多理由感到担忧，但一切都源于比简单的技术更深层次、更有影响的问题：在我们创造的过程中，是什么在激励着我们的心灵和思想？我相信，这个问题的答案也许比其他任何问题的答案都更能决定我们的未来。很多事情都取决于问题由谁来回答。随着人工智能领域逐渐变得更加多元、更加包容、对其他学科的专业知识更加开放，我也越来越有信心：我们能正确回答这个问题。”[9]当前，主流算法公平性标准（如

①8B尺寸达到GPT-4级性能！北大等提出医疗专家模型训练方法。详见https://www.36kr.com/p/2833371245775107。

群体平等、个体公平）大多源于自由主义传统，多与中国社会实际和文化价值存在脱节，如差异化的公平理念等。在中国文化中，“结果平等”与“机会平等”的权重与西方并不完全相同。西方文化侧重个体权利，中国文化关注人际关系网络中的平衡。在人工智能时代，开展科学技术与人文社会科学跨学科交叉性教育，对培养复合型人才十分关键，必须推动技术研发人员与人文社会科学学者之间的合作，确保技术发展兼顾社会影响和文化多元性。必须设计全面的人工智能伦理和法律框架，确保技术发展合规合理合法。要提高公众对人工智能的理解水平和参与程度，增强人工智能大模型的社会适应度。建立人工智能技术社会影响评估机制，及时发现其存在的问题，及时调整发展方向。人文社会科学学者需要学习和利用人工智能技术，但必须保持对人文核心价值的坚守，避免社会过度技术化。历史上，几乎每次技术革命都曾引发人文社会科学焦虑，但最终通过适应和创新实现了人文和社会科学学科的发展，我们坚信，这次人工智能革命也不会例外。不要怀疑人文社会科学的强大韧性和适应性，它能够在技术变革中找到自己的定位。在以上所有问题中，最为关键的是人工智能将由谁来掌握和使用。

（三）建立沙盒监管机制

人工智能企业需要在监管机构监督下测试高风险人工智能项目、产品（如自动驾驶等）。沙盒监管机构为企业划定一个“安全区”，允许它在一定期限内（比如6—24个月或者更长时间内）测试尚不完全成熟的人工智能技术，根据测试情况对其实施动态调整。沙盒管理的核心目标是降低创新门槛，避免由于传统审批流程阻碍技术快速迭代，同时通过小范围测试及时发现人工智能存在的技术、伦理、法律等问题，在测试数据基础上优化监管政策，避免“一刀切”。英国人工智能诊疗公司（Babylon Health）的人工智能诊断工具在通过英国国家医疗服务体系（National Health Service）的沙盒验证后正式上线。在北京、深圳等城市，通过数据跨境流动沙盒测试人工智能训练数据出境的安全性取得一定成效。本质上说，沙盒监管是以有限风险换取无限创新潜力的创新方式。沙盒监管不是放任自流，而是把实验室搬到监管机构的显微镜下，允许试错。未来，随着人工智能技术复杂度的提升，沙盒监管需在开放性与安全性之间找到更加精准的平衡点。

（四）建立人工智能负责人制度

建立和完善算法监管体系是各国面临的重要课题，它旨在确保算法透明、公平、安全，同时平衡技术创新与社会责任之间的关系。在大模型呈现主权化和文化多元化背景下，必须强调国家安全与社会稳定，在立法、备案和技术手段密切结合的基础上，构建全生命周期的算法治理体系。

作为一种治理制度，人工智能负责人制度（AI Accountability Officer或AI Governance Lead）是主要通过明确责任主体、规范开发流程和加强监督来防范人工智能风险的措施之一。人工智能负责人是指，在企业或组织中专门设置负责监督人工智能系统全生命周期合规性、安全性和伦理性的管理岗位，通常直接向企业董事会或CEO报告人工智能系统运行状况。我国的《生成式人工智能服务管理暂行办法》、欧盟的《人工智能法案》等都要求针对高风险人工智能系统建立明确的责任人制度。我国《互联网信息服务算法推荐管理规定》中明确规定，提供算法推荐服务应当遵守法律法规，尊重社会公德和伦理，遵守商业道德和职业道德，坚持主流价值导向，促进算法应用向上向善。[10]根据欧盟《人工智能法案》，高风险人工智能领域，如医疗、司法等必须确定责任人。欧盟《人工智能法案》把人工智能系统按风险进行分级，其核心逻辑是根据人工智能风险等级实施分级监管，定期更新分类标准，加强对人工智能系统的严格人工监督。美国微软设立“首席AI官”（Chief AI Officer），直接向CEO汇报人工智能运

行情况，统筹人工智能发展与安全。未来，随着人工智能渗透到各个关键领域，人工智能责任人制度需要进一步细化、规范，必须通过强化责任制度及其执行，实现创新与安全以及人机关系调适的双赢，最终找到人类与人工智能共存的边界。

（五）建立人工智能治理的公众参与机制

人工智能广泛应用涉及社会公共利益和全体人类福祉，其治理不能单纯依赖技术专家或专门治理机构，公众参与（Public Participation）必须成为确保人工智能发展符合人类价值观的关键机制。公众参与可以减少对“黑箱AI”的恐惧，如人脸识别滥用、社会焦虑等。欧盟《人工智能法案》要求，高风险人工智能系统必须寻求公众咨询，否则不得进入市场。公众监督会防止人工智能被某些利益集团操控，包括防止社交媒体算法操控选举等。由于员工参与和公众参与，谷歌才能确保其人工智能服务的真实需求，包括残障人士对使用无障碍人工智能的反馈和语音助手改进等。公众通过开源平台监督人工智能，如构建未来的人工智能社区（Hugging Face）的“模型卡”标注数据偏见等。推特（Twitter）用户集体测试GPT-3的偏见，推动OpenAI改进其内容过滤。应尝试通过社区培训提升公众监督能力，只有让公众成为人工智能的“共同设计者”而不是被动接受者，才能确保人工智能技术真正服务于人类整体利益。解决好技术开发人员常常忽视的社会接受度问题，确实需要跨学科合作、用户参与设计、社会影响评估、公众教育与沟通、政策与法规支持以及持续的意见建议反馈与改进等一系列环节。

提高人工智能社会参与水平的基础是提高全民人工智能素养，需要政府、教育机构、企业、媒体和社会组织协同推进。要从“人工智能扫盲”转向“全民人工智能思维”：把人工智能教育纳入教学培训体系。推进全社会人工智能普及；对在职人员开展人工智能技术培训，建立开放普惠型人工智能学习机制。建立“人工智能素养评估体系”。通过电影电视等文艺作品营造人工智能的文化氛围，使其融入日常生活。要重点关注老年群体等特殊群体的人工智能技能提升，布局好农村与偏远地区的人工智能基础设施。提升人工智能素养的本质是构建“人机共生”社会的基础设施和社会体制基础。在近期要解决“知不知道”，中期要实现“会不会用”，长期要培养“能不能创新性思考和提出创新性问题”。要发扬中国独特的优势，加强顶层设计，如以往推动的“全民数字素养行动”，不断提高社会动员能力，如基层治理和社会动员，探索出一条“政府主导—企业助力—全民参与”的中国特色人工智能社会治理之路，为全球人工智能治理提供中国方案。

（六）加强人工智能领域国际合作的体制机制

当前，人工智能的快速发展已超越国界，其潜在风险，诸如伦理冲突、军事化滥用、技术垄断等问题需要在全球协同层面上得以解决。国际合作必须成为确保人工智能安全的核心机制之一，其意义不仅在于风险防控，更关乎人类文明走向。一是尽管技术无国界，但人工智能风险全球化成为新时期风险管理的特点之一。一家美国公司开发的人工智能深度伪造（Deepfake）技术或许会被用于干扰他国选举。开源大模型，如美国互联网公司Meta的多语言大模型Llama被伊朗黑客用于生成恶意代码等。二是避免“逐底竞争”（Race to the Bottom）。如果各国为了争夺本国人工智能优势而放松监管，则会导致全球人工智能安全标准体系崩塌，如自动驾驶人工智能在不同国家的测试标准差异会引发交通事故等。部分国家为吸引人工智能企业，提供“监管洼地”，如允许不加限制的人脸识别数据采集，也会带来一系列意想不到的风险。三是数据协作。跨国医疗人工智能需多国患者数据训练，但必须合规共享。四是防止人工智能军备竞赛。自主武器系统的失控极有可能引发战争升级。联合国已经警告要预防“杀手机器人”的威胁。人工智能安全的本质是“集体行动

问题”——任何国家单边行动都无法应对全局性人工智能风险。人工智能可能是人类的最大机遇,也可能是最后的挑战,人类或许只有一次机会把它做对做好。最近,美国制定的《国家人工智能研发战略计划》,凸显了国际合作的重要性,这份报告提出通过跨学科合作来识别和减轻人工智能造成的伦理、法律问题,促进公平发展。该报告强调的重点是,可解释人工智能、隐私保护设计、模型偏见检测等关键技术,以及相应的政策框架。[11]国际合作需要超越短期利益博弈,在长期主义理念下确保人工智能服务于人类共同的福祉。

结语

人们总是在短期内高估了技术发明,而在长期上忽视了它的价值。尤其是在人工智能技术连连迭代的时候,长期主义尤为重要。如果在早期就认识了问题所在并加以跟踪,技术逐步成熟并深深嵌入体制机制后再进行治理会付出更少的治理代价。人工智能发展已经势不可挡,要开创人工智能时代的新型文明形态,就需要培育人文社会科学家与科学技术人员之间密切合作的文化,这种文化要求人文科学家深度了解科学技术发展的特点,科学技术人员深度了解人文社会科学的基本原理,人文科学家与自然科学家之间要达成默契——大家互相接纳、共同配合。尽管AI能模拟情感并提供更佳反馈,人类对真实社交连接的需求将持续存在,但这种基于生物本能的渴望无法被完全替代;AI不是直接取代人类工作,而是推动工作形式深刻变革,人类将成为决策者和创造性思维者,AI则作为工具增强人类能力①。还有,必须从现在起培育人类与机器之间默契配合的文化,这种文化是新的人类文明和智能文明时代的文化。随着具身智能的快速发展和融入社会生活,人类正站在人工智能时代的门槛上,有望达到前所未有的文明水平,新一代智能体,将主要通过经验学习并获得超越人类的能力。[12]人工智能时代的合作,不仅仅是科学技术人员的事情,也是人文社会科学人员的事情,更是全社会的事情。当全体社会成员把自己的未来和命运与人工智能技术开发结合起来时,大家才会产生一种自豪感、安全感,人类才会拥有更美好的未来,这应该也是新的文明形态所需要的核心价值。

①知识储备的价值正在让位于模式识别与综合能力。来自腾讯研究院AI速递,20250414。

[参 考 文 献]

[1]塞巴斯蒂安·洛塞.人工智能责任[M].上海人民出版社,2024:2.

[2][5][6][7][9]李飞飞.我看见的世界:李飞飞自传[M].中信出版社,2024:405,287,298,417-418,377.

[3]国际能源署预测:2030年全球数据中心电力需求将翻倍,人工智能成为最重要驱动力[EB/OL].中国能源网,https://www.cnenergynews.cn/news/2025/04/11/detail_20250411208837.html.

[4]阿尔伯特·温格.资本之后的世界:从资本稀缺到注意力稀缺,人类如何迎战经济奇点[M].中国财政经济出版社,2025:XXIII.

[8]平松类.AI时代的老年生活[M].浙江人民出版社,2024:26.

[10]国家互联网信息办公室等四部门发布《互联网信息服务算法推荐管理规定》[EB/OL].中国政府网,https://www.gov.cn/xinwen/2022-01/04/content_5666387.htm.

[11]National Science and Technology Council,Networking and Information Technology Research and Development Subcommittee[EB/OL].THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN,https://

www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.
[12]David Silver.Richard Sutton Welcome to the Era of Experience[EB/OL].https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf.

责任编辑:刘翠霞

Research on Potential Social Risks and Governance Institutional Mechanisms of Artificial Intelligence Technologies

Ding Yuanzhu

[**Abstract**] Artificial intelligence security and its risk governance, including the social governance of artificial intelligence, have become a topic of common concern to all walks of life at home and abroad. The main reasons for the risk of artificial intelligence include: artificial intelligence technology is still in the stage of exploration and development, cognitive misunderstandings caused by technological radicalism, imperfect complementary mechanism between science and technology and humanities and social sciences, "misalignment" between "technical capabilities" and "social expectations", and incomplete regulatory system. To build a controllable AI governance system and mechanism, we can focus on improving the evaluation connotation of artificial intelligence development, improving the interdisciplinary comprehensive research and development system and mechanism, establishing a sandbox supervision mechanism, establishing an artificial intelligence person system, establishing a public participation mechanism for artificial intelligence governance, and strengthening the institutional mechanism for international cooperation in the field of artificial intelligence.

[**Key Words**] Artificial Intelligence; Potential Risks; Governance Institutional Mechanisms