人工智能发展的前沿趋势、治理挑战与应对策略

薛 澜 王净宇

[摘 要] 人工智能发展已经进入大模型时代。在基础研发加速推进、产业应用加速落地的前沿趋势下，人工智能正负两方面的外部性不断释放，在产生巨大经济社会发展收益的同时，也从内生风险、应用风险、经济社会风险等多个维度对个体、国家乃至人类带来不容忽视的安全威胁。在此背景下，技术发展和治理不同步、监管者和被监管者信息不对称、风险防范成本和效益不对等、机制复合体治理不协调、地缘政治环境不稳定等人工智能治理面临的难点和挑战进一步暴露。对此，亟须从加大安全投入、健全监管制度、鼓励自我规制、加强国际治理合作等方面完善人工智能治理体系，更好应对人工智能快速发展带来的风险挑战。

[关键词] 人工智能；前沿趋势；风险挑战；治理路径

[中图分类号] D63 [文献标识码] A [文章编号] 1674-7453（2024）08-0004-10

2023年前后，以ChatGPT、Sora等为代表的大模型取得突破性发展，新一轮人工智能革命正在兴起、人工智能接近“技术奇点”的讨论空前激烈。[1]众多研究者认为，随着大模型时代通用人工智能逐渐可能甚至即将成为现实，人工智能能力和自主性的大幅提高将进一步放大其技术外部性，在给世界带来巨大机遇的同时，也带来了难以预知的各种风险和复杂挑战。但是目前，世界各国对前沿人工智能造成的潜在风险回应不足，相关的治理举措相对于人工智能技术的快速发展存在滞后性。[2]正因为如此，党的二十届三中全会审议通过的《中共中央关于进一步全面深化改革、推进中国式现代化的决定》强调，完善推动新一代信息技术、人工智能等战略性产业发展政策和治理体系，引导新兴产业健康有序发展，提出了“完善生成式人工智能发展和管理机制”“建立人工智能安全监管制度”两项具体要求。[3]

[基金项目] 中国科技创新2030重大专项“新一代人工智能风险防范与治理手段研究”（2023ZD0121700）。

[作者简介] 薛澜，清华大学苏世民书院院长，清华大学人工智能国际治理研究院院长，清华大学公共管理学院教授；王净宇，清华大学“水木学者”博士后，清华大学人工智能国际治理研究院助理研究员。

在此背景下，研究大模型等人工智能发展的前沿趋势、治理挑战与应对策略，对中国完善人工智能治理体系和监管制度、确保人工智能“智能向善”具有重要的理论和现实意义。本文基于“发展趋势—收益风险—治理挑战—治理对策”的逻辑思路，结合人工智能技术创新和产业应用的最新进展，分析前沿人工智能带来的发展收益、安全风险和治理难点，最终提出加强人工智能治理的对策建议。

一、人工智能发展的前沿趋势

自1956年美国达特茅斯会议正式提出人工智能概念以来，人工智能发展几经起落。直到21世纪人类进入互联网和大数据时代，计算机的计算和存储能力大幅提升，万物互联和传感技术的发展提供了丰富的数据资源，从大数据中自动获取知识的机器学习成为新一代人工智能的主要机制和技术驱动力，人工智能由此进入了新的快速发展期。[4]特别是近年来依靠先进算法、整合多模态大数据、汇聚大量算力的大模型取得重要突破，在掀起新一轮人工智能发展热潮的同时，不断推动人工智能迈向通用化的发展新阶段。有研究认为，迄今为止人工智能经历了两个最重要的发展阶段，其中之一就是当前以ChatGPT为最新标志，基于大数据的大模型时代。[5]在此背景下，人工智能发展进一步提速，呈现出基础研发加速推进、产业应用加速落地的整体趋势。

（一）人工智能基础研发加速推进

基础研究是科技创新的源头和先导。从研究文献、申请专利、开源项目等多个维度来看，人工智能领域的基础研发速度不断加快，奠定了当前人工智能发展热潮的重要基础。

第一，人工智能研究文献数量保持增长，机器学习异军突起。学术期刊、著作、会议论文等多种形式的文献发表数量是衡量不同领域基础研究发展情况的重要指标。根据斯坦福大学《2024年人工智能指数报告》，全球人工智能相关文献的年度发表数量自2010年起保持增长态势，2016年后更是进入了快速增长阶段，并于2021年和2022年达到了24万篇以上的峰值。在各细分领域中，机器学习相关文献发表数量于2016年后在人工智能研究中的占比快速上升，截至2022年已经占到人工智能发表总数的三分之一左右，构成了推动本轮人工智能发展高潮的核心力量。[6]

第二，人工智能专利数量激增，中国在全球人工智能创新中发挥重要作用。相较于文献数量，人工智能领域专利申请数量的快速增长期出现更晚，但爆发力更强。斯坦福大学《2024年人工智能指数报告》显示，全球人工智能领域专利申请和授予数量在2010至2017年间保持稳定，2018年后开始爆炸性增长。截至2022年，全球人工智能领域年度授权专利数量已经突破6.2万件，达到2018年的7倍以上。从国别比较的角度来看，中、美两国在人工智能专利数量上占据绝对优势，中国自2013年以来更是一直高居全球榜首，2022年授权专利数量占到全球总量的61.13%，为全球人工智能技术的发展贡献了重要力量。

第三，人工智能开源项目逐年递增，2023年迎来快速爆发。开源是指将源代码、设计文档或其他创作内容开放共享的技术开发和发行模式，历经40多年的发展，开源作为软件行业创新引擎的地位不断增强，已成为当前极为重要的科技创新渠道。[7]人工智能领域开源项目的开发和开源生态的完善极大地推动了全球人工智能的技术创新和产业发展。根据全球最大的开源社区GitHub统计，2011年至2023年的13年间，该平台上的人工智能开源项目由845个增长至181万个，增长超过2100倍。仅2023年一年，该平台上就新增了近70万个人工智能开源项目，点

赞数量更是突破 1220 万，达到了 2022 年的 3 倍以上。[8]

（二）人工智能产业应用加速落地

受益于人工智能基础研发的推进和技术范式的变革，产业界成为推动人工智能发展的主要力量，人工智能应用场景不断丰富、产业生态日趋成熟，产业应用的落地速度不断加快。

第一，产业界成为推动人工智能发展的主要力量，技术创新和产业应用的联系空前紧密。随着大模型成为引领本轮人工智能革命的技术范式，拥有更多数据和算力资源的产业界逐渐超越学术界成为推动人工智能发展的“主角”。根据 Bommasani（2023）等学者统计，在 2023 年全球发布的 149 个人工智能基础模型中，有 72.5% 来自产业界，仅有 18.8% 由学术界单独完成。[9]EPOCH 公司的统计进一步显示，产业界主导或参与训练的大模型在参数和算力使用等方面自 2014 年起开始逐渐超越学界，至 2023 年已经形成绝对优势。[10] 在此背景下，人工智能领域的产学研边界不断弱化，产业界主导的技术创新和产业应用正在充分互动、深度融合，人工智能从技术研发到实际应用的时间大大缩短。

第二，人工智能技术性能大幅提高，产业应用场景不断丰富。得益于大模型、多模态等技术的快速发展，人工智能系统的技术能力在近 5 年大幅提高。根据斯坦福大学《2024 年人工智能指数报告》，截至 2023 年底，人工智能已经在图像分类、基础阅读理解、自然语言推理、多语言理解、视觉推理等方面的标准化测试中超越或接近人类表现，仅在视觉常识推理、高级数学问题等方面暂时落后人类。随着人工智能的技术表现在越来越多的领域超越人类，其应用场景从传统单一化的数据分析转向更为广泛的内容创造，正在工业、医疗、金融、教育、电商、交通等诸多领域加速落地。

第三，人工智能商业模式日趋多元，产业生态逐渐成熟。商业模式关乎技术创新能否完成从价值创造到价值实现的闭环。目前，在人工智能应用场景不断丰富、赋能能力不断增强的背景下，以“模型即服务”（Model as a Service, MaaS）为主的人工智能产业生态逐渐成熟。在此模式下，人工智能供给侧正在出现大模型厂商打造通用大模型底座和行业大模型、各类企业或个人小模型充分竞争的基础业态。产业生态中基础层的大模型厂商、中间层的中小科技企业、应用层的传统企业各司其职，正在不断推动人工智能产业应用加速落地。

二、人工智能发展带来的收益与风险

在基础研发加速推进、产业应用加速落地的前沿趋势下，人工智能正负两方面的外部性不断释放，在产生巨大经济社会发展收益的同时，也带来了一系列安全风险。

（一）人工智能带来的发展收益

早在 2018 年，习近平总书记就在致世界人工智能大会的贺信中指出，新一代人工智能在全球范围内蓬勃兴起，为经济社会发展注入了新动能。[11] 人工智能作为重要的前沿引领技术和颠覆性技术，其发展不仅可以带来相关核心产业的巨大经济收益，还可以通过赋能千行万业催生新产业、新模式、新动能，助力解决可持续发展等人类共同面对的重要问题。

第一，人工智能核心产业发展带来巨大的经济收益。人工智能复杂的产业链环节和极高的技术附加值赋予了相关核心产业巨大的市场前景。以中国为例，工业和信息化部统计，截至 2023 年，中国人工智能核心产业规模已经突破 5000 亿元，企业数量超 4500 家。[12] 按照国务院《新一代人工智能发展规划》中的设计，到 2030 年，中国人工智能核心产业规模将超过 1 万亿元。[13] 在可见的将来，人工智能核心产业将在国民经济中扮演越来越重要的角色。

第二，人工智能发展赋能千行百业。人工智能是引领这一轮科技革命和产业变革的战略性技

术，具有溢出带动性很强的“头雁”效应。[14] 早在2017年，中国就在《新一代人工智能发展规划》中明确提出，人工智能将进一步释放历次科技革命和产业变革积蓄的巨大能量，并创造新的强大引擎，重构生产、分配、交换、消费等经济活动各环节，形成从宏观到微观各领域的智能化新需求，催生新技术、新产品、新产业、新业态、新模式，引发经济结构重大变革，深刻改变人类生产生活方式和思维模式，实现社会生产力的整体跃升。基于这一判断，《新一代人工智能发展规划》明确提出了到2025年推动新一代人工智能带动相关产业规模超过5万亿元、2030年带动相关产业规模超过10万亿元的发展目标。[15] 结合人工智能发展的前沿趋势来看，以大模型和生成式为代表的新型人工智能已经初步具备通用性、涌现性等特征，在智能交互、决策辅助、智能建模等方面不断取得突破性进展，赋能带动作用进一步加强。[16] 在此背景下，中国于2024年推动开展“人工智能+”行动，旨在促进人工智能与实体经济深度融合，以人工智能高质量发展和高水平应用培育经济发展新动能。受到技术创新加速和政策扶持力度加大的双重影响，人工智能带动的产业规模极有可能超过《新一代人工智能发展规划》设定的发展目标，对社会经济发展产生更为强大的推动作用。

第三，人工智能为可持续发展提供新的解决思路。可持续发展是破解当前全球性问题的“金钥匙”。[17] 人工智能带来的发展收益不只体现在经济效益方面，也体现在可助力人类解决可持续发展难题方面。根据《自然》杂志有关论文测算，在联合国提出的无贫穷、零饥饿、性别平等、清洁能源、应对气候变化等17个可持续发展目标的169个细分项目中，人工智能可以对其中134个产生积极作用，占比达到79%。例如，人工智能可以帮助人类整合多重形态的可再生能源，将电力需求与光能、风能相匹配，持续改善全球能源利用效率；通过大规模数据分析帮助人类设计最优化的环境保护行动方案；可参与气候变化影响的建模工作，增强人类应对气候变化的预警和治理能力……[18] 不仅仅是学术界，产业界、政府部门和非政府组织也都在关注人工智能在可持续发展问题上可能发挥的重要作用。中国在《全球人工智能治理倡议》中就专门提出，积极支持以人工智能助力可持续发展，应对气候变化、生物多样性保护等全球性挑战，充分阐释了人工智能协助人类推动可持续发展的巨大潜力。

（二）人工智能带来的安全风险

鉴于人工智能技术巨大的影响力和外部性，人工智能发展在带来各种发展收益的同时，也会带来难以忽视的安全风险。对此，有学者基于技术逻辑将人工智能带来的风险划分为“技术失控风险”“技术非正当使用风险”和“技术社会效应风险”三类。[19] 也有学者从分类治理的视角出发，将人工智能风险划分为人工智能技术、产品或服务本身的“技术风险”和人工智能应用过程中的关联系统、环境及行为主体所产生的“业态风险”。[20] 虽然不同领域的研究者基于各自研究视角对人工智能风险的分类方式有所不同，但总体上基本认同人工智能发展将从技术本身、技术使用、社会影响各个维度对个体、国家乃至人类带来不容忽视的安全威胁。

第一，人工智能技术本身可能产生技术幻觉、算法歧视甚至技术失控等内生风险。其一，人工智能产生不准确或误导性输出的技术幻觉风险。既有研究认为，当前的大模型尚不具备判断真假的机制和“知不知”的能力，其内容输出更接近于“重构资讯”，在不可能穷尽所有代表性数据的情况下，无法确保输出内容的真实性。[21] 加之训练数据过度拟合、上下文理解有限、模型架构缺陷等技术问题的影响，人工智能在实际应用过程中可能输出错误信息的技术幻觉问题很难在短时间内得到解决，从而对内容真实性要求较高的应用场景构成风险挑战。其二，人工智能的算法歧视风险。目前，以深度学习等算法为代表的机器学习是人工智能领域主

流的技术范式。这种算法基于“大数据集”进行自我训练、自我学习最终形成“规则集”的过程，实质上是对于过往人类社会模式特征的总结并将其用于对未来社会的感知与决策，不可避免地会复制并延续当前社会的既有格局与特征。[22] 受制于人和数据天然偏见的结构性不平等，人工智能的技术发展阶段和发展路径决定了其在歧视面前近乎不设防的虚弱状态。[23] 人工智能难以消除的算法歧视又会进一步放大使用者的偏见和冲突。其三，人工智能技术失控风险。以大模型、生成式为代表的新型人工智能需要通过机器学习特别是神经网络技术进行训练，相关技术透明性和可解释性不足的特性导致人工智能发生和涌现产生的机理目前仍是“黑箱”，[24] 无论是用户还是研发者都不能完全控制大模型的行为。在此背景下，一旦产生自我保护、自我发展意识的未来人工智能系统脱离人类控制，很有可能对人类生存构成威胁。

第二，人工智能技术的大规模普及可能产生误用、滥用、恶用等应用风险。人工智能作为一种重要的通用性、基础性工具，一旦遭到误用、滥用，极有可能造成数据泄露、深度伪造、恐怖主义、技术军事化等风险。其一，数据隐私和信息安全风险。人工智能海量的训练数据往往包含用户的隐私信息，在办公、医疗、教育、交通等领域的垂直应用更是涉及大量敏感数据，如果相关信息被错误或恶意调用，将对信息安全构成严重威胁。其二，深度伪造和低成本造假风险。当前的人工智能技术可以利用多种神经网络模型实现文本、图像、音频、视频等内容的篡改、伪造和自动生成，产生以假乱真的效果。虽然这种深度伪造技术在教育、娱乐等领域释放了巨大的应用潜力，但也显著降低了造假成本，导致利用深度伪造技术仿冒他人身份实施违法犯罪行为的现象越来越常见。[25] 其三，恐怖主义风险。人工智能技术的发展和普及客观上增强了恐怖组织在不对称冲突中的实力、降低了恐怖分子实施恐怖袭击的成本，恐怖组织和极端势力很可能利用人工智能技术扩大行动范围，甚至更新恐怖主义的形态。[26] 其四，技术军事化风险。人工智能是军事变革的潜在推动者和力量倍增器，以致命性自主武器系统为代表的军用人工智能技术的发展将严重冲击既有的国际军备控制体系，引发新一轮全球性军备竞赛，从而对全球战略稳定构成严峻挑战。[27]

第三，人工智能技术变革的颠覆性影响引发劳动力结构调整、社会不平等加剧等经济社会风险。其一，对劳动力市场造成冲击。有研究认为，人工智能不同于以往历次科技革命和产业变革中的技术创新，其发展不仅可以拓展人类的生产能力、取代简单重复性的常规工作，还可以通过学习执行越来越多的认知性工作，在一定程度上超越人类改造世界的工具和手段的范畴，具有以全新方式替代人类劳动的潜质。[28] 相关研究表明，在人工智能通过替代效应、创造效应和生产率效应影响劳动力需求和工资水平的过程中，替代效应暂时占据主导地位。其二，加剧社会不平等。从市场结构的角度看，人工智能通过改变市场与行业结构降低了劳动收入份额；从技能结构的角度看，人工智能通过推动就业极化拉大了高收入群体和中低收入群体之间的收入差距。总体来看，人工智能技术进步具有扩大收入不平等的张力，在缺少有效公共政策的情况下将导致就业率下降、就业与工资极化、收入与财富不平等加剧等一系列问题，从而造成严重的社会不公平现象。[29] 除此之外，还有研究者认为，人工智能将改变人类社会的发展进程，重新建构和形塑人与自然、人与技术以及个体与国家、个体与社会的关系，[30] 在推动社会形态演变的过程中还有可能对经济安全和社会稳定产生一系列目前难以预知的冲突与风险。

三、人工智能治理的难点与挑战

人工智能的快速发展蕴含巨大的发展机遇，也

涉及一系列不容忽视的安全风险。对此，《新一代人工智能发展规划》提出，要在大力发展人工智能的同时高度重视可能带来的安全风险挑战，确保人工智能安全、可靠、可控发展，为人工智能治理描绘了高标准、前瞻性的治理蓝图。《中共中央关于进一步全面深化改革、推进中国式现代化的决定》在进一步全面深化改革的原则中同样强调，坚持系统观念，处理好发展和安全等重大关系。[31]综合人工智能发展的前沿趋势及其产生的收益风险来看，目前主要存在以下五个方面的治理难点与挑战。

（一）技术发展和技术治理不同步

在人工智能基础研发、产业应用共同加速的背景下，人工智能发展同时具有快速迭代和不可预测的特征，对治理系统与技术系统的同步演进提出了巨大挑战。

其一，人工智能技术的迭代速度超出人类预期，现有治理系统的演进速度难以追上技术系统的发展步伐。在技术创新规律方面，摩尔定律是预测计算机处理能力的经典标准，但人工智能尤其是通用人工智能系统的进化和迭代速度远远超出这一定律。在此背景下，如何克服社会、经济、法律等治理系统演进的高昂成本，根据人工智能的发展情况及时完成治理体系的快速反应、协同共振，是人工智能治理的重要挑战。

其二，人工智能技术的智能涌现具有不可预测性，治理系统很难准确预判人工智能的发展及其复杂的经济社会影响。一方面，与传统技术变革相比，以大模型为代表的人工智能初步具备了自我创造、超强学习、超级进化的特性，这一技术潜力导致人工智能技术的发展方向和突破节点难以预测。另一方面，未来通用人工智能、世界模型、实体智能等技术的发展可能引发社会经济结构整体性、颠覆性变革，治理系统很难及时识别人工智能发展潜在的、难以预知的冲突和风险，并提前针对相关风险做好准备。

（二）监管者和被监管者信息不对称

在当前人工智能发展的前沿趋势中，技术创新和产业应用的联系空前紧密，产业界已经成为推动人工智能发展的主要力量，监管者和被监管者之间不断扩大的信息盲区构成了人工智能治理的又一难点。

其一，监管者很难准确掌握技术发展的实时动态。一般来说，政府作为监管主体，在识别治理对象及其风险、权衡治理目标、选择治理工具时需要掌握技术创新和应用的最新动态；而被监管者作为一线创新者和应用者，对技术发展和技术风险问题具有天然的信息优势，双方需要通过合适的“委托—代理”结构设计来缓解信息不对称问题。[32]但是人工智能创新应用内在逻辑的复杂性“黑箱”进一步扩大了双方的信息差距，甚至有可能造成“共同无知”的新局面，从而放大了监管者和被监管者之间信息沟通、合作治理的难度。

其二，被监管者很难明确监管者的治理目标。传统技术系统自身的运作逻辑、功能性能、应用场景比较清晰，引致风险的原因及需要实现的治理目标往往也较为明确。但是人工智能发展的快速迭代和不可预测性导致监管者需要不断探索、权衡、调整治理目标和治理机制。在此背景下，被监管者需要及时了解政府规制关切的最新变化，适应人工智能时代治理要求的难度也在加大。

（三）风险防范成本和效益不对等

人工智能基础性、通用性的技术特征在释放巨大发展潜力的同时大幅提高了风险防范成本，如何克服风险防范成本和效益的不对等，寻求发展和安全之间的最佳平衡，也是人工智能治理的难点所在。

正如前文所述，人工智能对人类社会的影响具有颠覆性，如果遭到误用、滥用、恶用，可能引发社会、经济、军事等领域的重大安全威胁，一旦脱离人类控制甚至还有可能对人类生存构成威胁。如果不加以监管和治理，这些切实存在的安全风险

可能造成不可逆的严重后果。但另一方面，不同于需要大量基础设施投入、物理可见性较强的传统公共危险源，人工智能的通用性和易得性使其潜在风险几乎无处不在，极大提高了风险防范的社会成本。要全面防范人工智能带来的安全风险需要耗费大量的公共资源，甚至以部分牺牲人工智能发展为代价。如何在高昂的风险防范成本、巨大的潜在风险损失、不断提升的风险概率之间权衡，是监管者必须面对的重要考验。

（四）机制复合体治理不协调

人工智能治理是涉及法律法规、行业标准、国际协调的复杂系统，需要不同国家、国际组织、相关企业的共同参与。如何在机制复合体的复杂框架下实现国际治理协调，是人工智能治理亟待解决的重要问题。

理论上，机制复合体是指在一定问题领域或一定地理范围内多种国际规则或国际行为体同时起到规范或治理作用的机制集合，具有治理主体多元、治理层次丰富、治理机制灵活、治理手段多样等特征。[33] 目前，以联合国为代表的国际组织、以七国集团为代表的多边合作机制、以中美为代表的人工智能大国、以微软和谷歌等为代表的跨国企业等各类国际行为体都在深度参与人工智能治理。这些参与主体在治理理念、关注议题、治理偏好、治理能力上存在差异，导致人工智能国际治理呈现嵌套、重叠、平行的碎片化局面，存在合理性、公平性、有效性等一系列治理赤字。[34] 在此背景下，如何协调这些相互重叠、不同层次、不同诉求的参与主体，建立受到广泛认可的人工智能全球治理机制复合体，是人工智能治理的重要挑战。

（五）地缘政治环境不稳定

人工智能作为颠覆性的前沿技术，是大国博弈的关键领域。在当前大国战略竞争的地缘政治环境下，人工智能治理问题将变得更加复杂。

一方面，人工智能国际治理需要掌握前沿技术的大国建立沟通协调机制，通过信息交换和自我控制在国际治理体系中发挥更为积极的作用。[35] 但另一方面，人工智能拥有从经济、军事等多个维度影响国家力量对比，甚至引发新一轮大国兴衰的巨大潜力。[36] 大国战略竞争的地缘政治环境导致大国协调变得更加困难，部分国家甚至以意识形态划线或构建排他性集团，恶意阻挠他国人工智能发展，利用技术垄断和单边强制措施制造发展壁垒，恶意阻断全球人工智能供应链，不但阻碍全球人工智能发展，对人工智能治理也产生了负面影响。

四、人工智能的治理建议

2018 年至今，习近平总书记多次强调要重视人工智能发展，营造创新生态，重视风险防范，[37] 为中国推动人工智能创新发展和风险防范指明了方向。从 2017 年《新一代人工智能发展规划》提出“加强前瞻预防与约束引导，最大限度降低风险，确保人工智能安全、可靠、可控发展”的战略要求，到 2019 年国家人工智能治理专家委员会推出 8 项具体的人工智能治理原则，再到 2020 年后一系列法律法规、管理办法、指导意见、国际倡议的密集发布，中国的人工智能治理机制不断完善，具有多元主体、多维共治、多种工具、敏捷协同特征的中国特色人工智能治理体系初见雏形。2024 年 7 月，《中共中央关于进一步全面深化改革、推进中国式现代化的决定》提出“完善生成式人工智能发展和管理机制”“建立人工智能安全监管制度”两项具体要求，进一步明确了下一阶段人工智能的发展和治理方向。[38] 结合人工智能发展的前沿趋势、收益风险、治理难点来看，中国可以在当前治理机制的基础上，从以下几个方面进一步完善人工智能治理体系、建立人工智能安全监管制度，更好应对人工智能快速发展带来的风险挑战。

第一，加大人工智能安全领域的研发投入，加快人工智能安全治理能力建设。为了确保人工智能系统安全、可靠、可控，要对人工智能模型进行严

格的能力评估、对齐训练和红线测试，并对各类潜在风险进行及时的评估、追踪、识别、防范和响应，这些工作都需要大量的技术储备和研发投入。但是目前，产业界和学术界过多关注提高人工智能系统能力的科技创新，对于人工智能安全和治理领域的研究相对不足。针对这一现象，包括三位图灵奖得主、一位诺贝尔奖得主及姚期智、薛澜等中国学者在内的15位专家在《科学》（Science）杂志联名发声，呼吁至少将研发预算的三分之一投入人工智能安全研究。[39]对此，中国可以在建立未来产业投入增长机制的过程中通过提高研发费用加计扣除比例等方式鼓励和要求大型科技公司在生成式等前沿人工智能技术开发中围绕风险防范问题提高研发投入；在建立科技发展、国家战略需求牵引的学科设置调整机制和人才培养模式中重点加强人工智能安全和治理领域的人才培养，为人工智能安全监管和稳定发展提供人力基础。

第二，建立审慎包容、分级分类的人工智能安全监管制度，通过敏捷治理摆脱治理困境。如前所述，党的二十届三中全会将确保人工智能安全列入推进国家安全体系和能力现代化的重要工作。鉴于人工智能的发展趋势和技术特性，当下很难确定其演变、用途、风险、回报，因而难以提前明确监管对象和监管内容，相关治理只能在监管互动中前行。[40]在此过程中，政府监管部门可以采用敏捷治理的基本理念，推动人工智能技术发展的利益相关方形成持续性的共同学习和探索机制，根据前沿人工发展及其带来的收益风险及时调整治理对象和内容，采用分级分类、监管沙盒等方式，通过治理结构、治理策略、治理工具的灵活创新克服人工智能治理和监管的难点挑战。

第三，重视企业科技创新主体地位，鼓励行业和企业层面的自我规制。面对人工智能技术的快速迭代，政府部门一方面可以通过制定随模型能力变强而更加严格的安全标准，要求企业对人工智能系统造成的可预防伤害承担责任等措施加强外部规制。另一方面，产业界作为当前人工智能科技创新最重要的推动力量，也需要在人工智能的治理和监管框架中发挥重要作用。对此，政府部门可以在外部监督的基础上，激励引导行业组织和经营主体发挥自律作用，通过行业标准、企业规范等被监管者内部的自我规制缓解人工智能领域日益突出的发展治理同步性问题和外部规制信息不对称问题。

第四，依托多边平台，参与、引领人工智能全球治理。人工智能治理具有极强的外部性，攸关全人类命运，是世界各国亟须共同应对的重要挑战，积极参与人工智能全球治理是中国引导人工智能健康有序发展的基础和前提。面对协调机制复合体的治理难题，中国可以重点选择在联合国框架内加强人工智能治理协作和标准塑造，以能力建设为抓手推动构建开放、公正、有效、具有广泛共识的人工智能治理框架和标准规范。

第五，完善双边对话机制，加强与主要人工智能大国的政策沟通。作为引领全球人工智能发展和治理的主要大国，主要人工智能大国的竞合关系不仅事关双边关系大局，还会对全球人工智能治理成效甚至人类命运产生重要影响。面对大国竞争等地缘政治因素干扰，中国可以基于中美元首旧金山会晤共识，在保持中美政府间人工智能对话的同时，积极推动中美科学界、产业界之间的二轨对话，围绕设置安全红线、防范极端风险、推进全球治理等双方共同关切的问题加强对话、管控分歧、拓展合作，努力在大国战略竞争的背景下实现人工智能发展和治理领域的大国协调。

[参考文献]

[1]韩永辉,张帆,彭嘉成.秩序重构:人工智能冲击下的全球经济治理[J].世界经济与政治,2023(1).

[2]Bengio Y,Yao A,Xue L,et al.Managing Extreme AI Risks Amid Rapid Progress[J].Science,2024(5).

[3][31][38]中共中央关于进一步全面深化改革 推进中国式现代化的决定[EB/OL].中国政府网,https://www.gov.cn/

zhengce/202407/content_6963770.htm.
[4]郭毅可.论人工智能历史、现状与未来发展战略[J].人民论坛·学术前沿,2021(23).
[5]王天恩.人工智能通用化及其实现途径[J].中国社会科学,2024(3).
[6]Nestor Maslej et al.The AI Index 2024 Annual Report[EB/OL].AI Index Steering Committee,Institute for Human-Centered AI,https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.
[7]隆云滔,王晓明,顾荣,包云岗.国际开源发展经验及其对我国开源创新体系建设的启示[J].中国科学院院刊,2021(12).
[8]Nestor Maslejet al.The AI Index 2024 Annual Report[EB/OL].AI Index Steering Committee,Institute for Human-Centered AI,https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.
[9]Bommasani et al. The Foundation Model Transparency Index[J/OL].ArXiv Preprint ArXiv:2310.12941,2023(10).
[10]Parameter, Compute and Data Trends in Machine Learning[EB/OL].Epoch AI,https://epochai.org/blog/announcing-updated-pcd-database.
[11]习近平致信祝贺二〇一八世界人工智能大会开幕强调 共享数字经济发展机遇 共同推动人工智能造福人类[N].人民日报,2018-09-18.
[12]赛迪智库人工智能产业形势分析课题组.2024年我国人工智能产业展望[J].软件和集成电路,2024(1).
[13][15]国务院关于印发新一代人工智能发展规划的通知(国发〔2017〕35号)[EB/OL].中国政府网,https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
[14]习近平在中共中央政治局第九次集体学习时强调 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.
[16]段雨晨.以人工智能赋能高质量发展[J].红旗文稿,2024(7).
[17]习近平.坚持可持续发展 共创繁荣美好世界——在第二十三届圣彼得堡国际经济论坛全会上的致辞[N].人民日报,2019-06-07.
[18]Vinuesa R,Azizpour Het al.The Role of Artificial Intelligence in Achieving the Sustainable Development Goals[J].Nature Communications,2020(1).
[19]陈小平.人工智能:技术条件、风险分析和创新模式升级[J].科学与社会,2021(2).
[20]薛澜,贾开,赵静.人工智能敏捷治理实践:分类监管思路与政策工具箱构建[J].中国行政管理,2024(3).
[21]陈小平.大模型关联度预测的形式化和语义解释研究[J].智能系统学报,2023(4).
[22]贾开.人工智能与算法治理研究[J].中国行政管理,2019(1).
[23]李成.人工智能歧视的法律治理[J].中国法学,2021(2).
[24]张海柱.行动者网络理论视域下的算法黑箱与风险治理[J].科学学研究,2023(9).
[25]龙俊,王天禹.人工智能深度伪造技术的法律风险防控[J].行政管理改革,2024(3).
[26]谢磊.人工智能时代的恐怖主义:挑战与应对[J].和平与发展,2021(2).
[27]张煌,杜雁芸.人工智能军事化发展态势及其安全影响[J].外交评论(外交学院学报),2022(3).
[28]李磊,王小霞,包群.机器人的就业效应:机制与中国经验[J].管理世界,2021(9).
[29]陈斌开,徐翔.人工智能与社会公平:国际经验、影响机制与公共政策[J].国际经济评论,2024(3).
[30]侯利文,李亚璇.人工智能发展限度与社会可能[J].科技进步与对策,2021(9).
[32]贾开,赵静,傅宏宇.应对不确定性挑战:算法敏捷治理的理论界定[J].图书情报知识,2023(1).

[33]KarenAlter,Kal Raustiala.The Rise of International Regime Complexity[J].Annual Review of Law and Social Science,2018(1).

[34]贾开,俞晗之,薛澜.人工智能全球治理新阶段的特征、赤字与改革方向[J].国际论坛,2024(3).

[35]高奇琦.从大国协调到全球性机制:人工智能大模型全球治理路径探析[J].当代世界,2024(5).

[36]傅莹.人工智能对国际关系的影响初析[J].国际政治科学,2019(1).

[37]中共中央政治局召开会议 分析研究当前经济形势和经济工作 中共中央总书记习近平主持会议[EB/OL].中国政府网,https://www.gov.cn/yaowen/2023-04/28/content_5753652.htm.

[39]Bengio Y,Yao A,Xue Let al.Managing Extreme AI Risks Amid Rapid Progress[J].Science,2024(5).

[40]薛澜,赵静.人工智能国际治理:基于技术特性与议题属性的分析[J].国际经济评论,2024(3).

责任编辑：刘翠霞

Cutting-edge Trends in AI Development, Governance Challenges, and Response Strategies

Xue Lan & Wang Jingyu

[**Abstract**] The advent of large language models (LLMs) has ushered in a new era of AI development, characterized by accelerated advancements in foundational research, paradigm shifts in technological approaches, and rapid adoption across industries. The rapid advancement of AI has amplified both the positive and negative externalities, contributing significantly to economic and social progress while also posing security threats at personal, national, and global levels. The challenges in AI governance, such as the gap of technological and governance development,information asymmetries between regulators and the regulated, imbalance of costs and benefits in risk management, lack of coordination within the international regime complex, and instability in geopolitical environments have intensified. Faced with these challenges,China needs to strengthen its AI governance framework by increasing security investments, adopting agile governance practices, promoting self-regulation, and enhancing international governance cooperation, aiming to better tackle the risk challenges arising from the rapid advancement of artificial intelligence.

[**Key Words**] AI; Technology Trends; Challenges of Risks; Paths to Governance