

国际人工智能伦理治理行动计划

人工智能是引领新一轮科技革命和产业变革的战略性技术，为人类社会开辟了新前景，同时带来不容忽视的科技伦理挑战。为增进人类福祉，促进人类社会和平发展和可持续发展，推动构建包容、安全、可信、可持续的人工智能未来，我们提出《国际人工智能伦理治理行动计划》，呼吁各方共同行动，凝聚伦理治理共识，充分落实联合国《未来契约》及其附件《全球数字契约》有关承诺，为全人类创造技术创新成果普惠共享、伦理治理机制透明可信、开放全面、发展可期的智能未来。

我们重申，人工智能的开发和应用必须以国际法为基础，包括《联合国宪章》，并必须尊重、保护和促进人权与基本自由。人工智能系统在其全生命周期内的设计、部署和使用，应维护隐私权与数据保护，确保人类监督与问责，并与《2030 年可持续发展议程》保持一致。应特别关注弱势群体的权益，包括妇女、儿童、老年人、残障人士、工人、青年、中小微企业以及发展中国家的企业。

一、推动各方广泛参与伦理治理

坚持共商共建共享的全球治理观，呼吁形成政府引领、国际组织交流互信、企业与科研机构创新发展、民间机构与公民主动的伦理治理格局。倡导建立跨国家、跨区域、跨行业、跨学科的人工智能伦理治理协作机制。支持联合国在人工智能全球治理中发挥主渠道作用，增强发展中国家在人工智能全球治理中的代表性和发言权，共同维护平等发展和利用人工智能的权利。

二、强化人工智能全生命周期伦理治理

倡导人工智能产业主体关注伦理风险，推动伦理治理从事后防控向研发源头延伸，引导产业主体坚持公平性和非歧视性原则，将伦理治理纳入人工智能科学研究、技术开发、产品部署与运营维护的产业全生命周期。在数据采集与算法设计环节，提升数据真实性和准确性，提升人工智能系统的透明性与可解释性。在人工智能系统部署与使用环节，避免技术的误用、滥用与恶用。

三、深化人工智能伦理风险防控合作

鼓励各国结合人工智能系统技术特征、应用场景、影响范围和潜在后果，探索建立伦理风险分类分级识别机制。关注前沿人工智能系统可能引发的重大伦理风险，探索建立面向关键信息基础设施、重要行业和社会公众的分级风险告知机制。共同探索开展伦理风险分级分类定期评估，适时更新伦理风险分级分类参考标准。呼吁各国重点关注智能体、具身智能等新型人工智能形态，围绕数据敏感程度、工具调用权限、自主决策能力等因素，形成差异化风险处置规则。鼓励开源伦理社区加强伦理治理，防范开源伦理风险。

四、形成敏捷治理的动态伦理治理机制

积极适应人工智能系统快速迭代和智能体自主行动能力增强的趋势，实施敏捷治理。鼓励建立人工智能伦理风险案例库和重大伦理风险事件信息交流机制，适时调整案例库，动态优化交流机制，为敏捷治理提供实证基础。呼吁在充分尊重各国主权、安全关切、法律法规的前提下，基于共同商定的标准与程序，自愿分享风险警示、处置经验和治理工具，共同提升各国敏捷治理能力。

五、培育产业链协同的伦理治理生态

呼吁在加强治理的同时注重服务供给，培育覆盖研究、开发、应用、推广全链条的全球人工智能伦理治理生态。鼓励各国政府明确基础模型提供者、智能体开发者、部署者、工具和接口提供者、终端厂商、分发平台、运维主体和使用者的权责边界。鼓励开发可解释性、鲁棒性、公平性、可靠性、隐私保护、偏见矫正等伦理治理技术工具，推进人工智能科技伦理数据集共建共享和治理技术工具开源。

六、营造伦理治理社会氛围

开展人工智能伦理治理能力提升国际合作行动，重点支持发展中国家伦理治理能力建设。呼吁加强科技伦理治理人才培养与国民教育，将科技伦理纳入高等教育课程体系，培养兼具技术能力与伦理素养的复合型人才。鼓励面向全民搭建伦理治理宣传教育平台，发挥社会团体与大众传媒作用，共同推动公众伦理意识提升与行业自律文化建设。尊重、保护和促进妇女、儿童、老年人、残障人士等群体的权益，努力弥合数字鸿沟。

人工智能的未来由全人类共同塑造，科技伦理治理的道路需要各方携手同行。我们倡议将伦理先行确立为治理理念，将伦理自觉内化为行动追求，努力构建更加安全、可信、公平、包容的全球人工智能伦理治理体系，促进人工智能技术造福全人类，推动构建包容、安全、可信、可持续的人工智能未来。