新一代人工智能治理原则——发展负责任的人工智能

6月17日，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》（以下简称《治理原则》），提出了人工智能治理的框架和行动指南。

近年来，人工智能迅速发展，正在深刻改变人类社会生活、改变世界。为促进新一代人工智能健康发展，加强人工智能法律、伦理、社会问题研究，积极推动人工智能全球治理，新一代人工智能发展规划推进办公室成立了国家新一代人工智能治理专业委员会。

起草《治理原则》是委员会今年的重点工作，《治理原则》经过网上建议征集、专家反复研讨、多方征求意见等环节，凝聚了广泛共识。

《治理原则》旨在更好协调人工智能发展与治理的关系，确保人工智能安全可控可靠，推动经济、社会及生态可持续发展，共建人类命运共同体。《治理原则》突出了发展负责任的人工智能这一主题，强调了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八条原则。

《治理原则》全文如下：

新一代人工智能治理原则——发展负责任的人工智能

全球人工智能发展进入新阶段，呈现出跨界融合、人机协同、群智开放等新特征，正在深刻改变人类社会生活、改变世界。为促进新一代人工智能健康发展，更好协调发展与治理的关系，确保人工智能安全可靠可控，推动经济、社会及生态可持续发展，共建人类命运共同体，人工智能发展相关各方应遵循以下原则：

一、和谐友好。人工智能发展应以增进人类共同福祉为目标；应符合人类的价值观和伦理道德，促进人机和谐，服务人类文明进步；应以保障社会安全、尊重人类权益为前提，避免误用，禁止滥用、恶用。

二、公平公正。人工智能发展应促进公平公正，保障利益相关者的权益，促进机会均等。通过持续提高技术水平、改善管理方式，在数据获取、算法设计、技术开发、产品研发和应用过程中消除偏见和歧视。

三、包容共享。人工智能应促进绿色发展，符合环境友好、资源节约的要求；应促进协调发展，推动各行各业转型升级，缩小区域差距；应促进包容发展，加强人工智能教育及科普，提升弱势群体适应性，努力消除数字鸿沟；应促进共享发展，避免数据与平台垄断，鼓励开放有序竞争。

四、尊重隐私。人工智能发展应尊重和保护个人隐私，充分保障个人的知情权和选择权。在个人信息的收集、存储、处理、使用各环节应设置边界，建立规范。完善个人数据授权撤销机制，反对任何窃取、篡改、泄露和其他非法收集利用个人信息的行为。

五、安全可控。人工智能系统应不断提升透明性、可解释性、可靠性、可控性，逐步实现可审核、可监督、可追溯、可信赖。高度关注人工智能系统的安全，提高人工智能鲁棒性及抗干扰性，形成人工智能安全评估和管控能力。

六、共担责任。人工智能研发者、使用者及其他相关方应具有高度的社会责任感和自律意识，严格遵守法律法规、伦理道德和标准规范。建立人工智能问责机制，明确研发者、使用者和受用者等的责任。人工智能应用过程中应确保人类知情权，告知可能产生的风险和影响。防范利用人工智能进行非法活动。

七、开放协作。鼓励跨学科、跨领域、跨地区、跨国界的交流合作，推动国际组织、政府部门、科研机构、教育机构、企业、社会组织、公众在人工智能发展与治理中的协调互动。开展国际对话与合作，在充分尊重各国人工智能治理原则和实践的前提下，推动形成具有广泛共识的国际人工智能治理框架和标准规范。

八、敏捷治理。尊重人工智能发展规律，在推动人工智能创新发展、有序发展的同时，及时发现和解决可能引发的风险。不断提升智能化技术手段，优化管理机制，完善治理体系，推动治理原则贯穿人工智能产品和服务的全生命周期。对未来更高级人工智能的潜在风险持续开展研究和预判，确保人工智能始终朝着有利于人类的方向发展。

国家新一代人工智能治理专业委员会

2019年6月17日