CAICT 中国信通院

人工智能与经济社会研究中心
Research Center of AI Economy and Society

全球人工智能治理体系报告

（2020）

中国信息通信研究院政策与经济研究所
人工智能与经济社会研究中心

2020 年 12 月

版权声明

前　言

人工智能作为引领新一轮科技革命和产业变革的战略性技术，正在重塑生产方式、优化产业结构、提升生产效率、赋能千行百业，推动经济社会各领域向智能化加速跃升。但“每个硬币都有正反面”，人工智能在为人类生产生活带来诸多便利的同时，也在触发如泄露个人隐私、冲击就业格局、危害公共安全等风险问题，给社会治理带来全新挑战。人工智能治理这一概念应运而生，推动构建全球人工智能治理体系正在成为各方共同诉求，国际组织积极探索全球性伦理原则并推动各方达成共识，各国政府基于各自产业发展特点采取不同的治理思路及举措，行业组织着力构建全面、有效的人工智能标准规范体系，科技企业则更加关注如何将伦理原则落地践行于自身产品及服务之中。

为更好地认识和推动人工智能治理实践，加快构建适宜人工智能产业发展的治理体系，中国信息通信研究院政策与经济研究所研究团队开展研究并形成本报告。主要包含以下内容：第一部分分析了人工智能对经济社会发展带来的促进作用及风险挑战，剖析了人工智能治理的痛点及难点；第二部分梳理总结了全球部分国家及地区的人工智能治理进展及特点；第三部分从治理目标、治理主体、治理手段、治理评价等维度，提出了构建人工智能治理体系的设计思路；第四部分提出了持续完善我国人工智能治理体系的几点思考。期待本报告能够为社会各界提供有价值的参考，不妥之处请不吝指正。

目 录

一、人工智能带来全球治理新挑战

当前，新一轮科技革命和产业变革正在深入发展，人工智能作为有望引领未来变革的战略性信息技术，在实现快速发展、广泛应用的同时，也带来了安全、隐私、公平等诸多新问题新挑战。全球各国纷纷呼吁加强人工智能治理，营造规范、健康、可持续的发展环境。

（一）人工智能引领经济社会智能化变革

人工智能至今历经三次发展浪潮，当前已成为科技创新、经济发展、社会进步、民生改善的重要驱动力量。人工智能是研究类人机器或系统的科学技术。这类机器或系统在某种程度上具有可替代人的智力或行为能力，其发展及应用将对人类的经济社会发展产生深远影响。人工智能一词源于1956年召开的美国达特茅斯会议，在60多年间经历了三次发展浪潮。在前两次浪潮中，由于人工智能的商业化价值和社会关注程度较低，国家与社会投入有限，加之理论技术未成熟、数据及算力等基础支撑不足、应用场景受限等问题，并未达到预期发展效果。随着新兴信息技术的快速发展，特别是机器学习算法的革新、海量数据的积累、计算能力的提升和新基建的演进，全球迎来了新一轮人工智能发展浪潮。当前，以深度学习、跨界融合、人机协同、群智开放、自主操控为特征的新一代人工智能技术持续创新突破，在促进科技创新、培育新兴产业、改造升级传统业态、加快实体经济转型等方面的作用日益突显。

人工智能将促进新工具与新技术的研发，提升社会生产效率。人类劳动经过手工劳作、工具使用、动力转换三大阶段，即将进入智能

替代的全新生产时代。第一次生产力升级是专用生产工具的发明，将人类生产效率大幅度提升。第二次生产力升级拉开了自动化的序幕，利用化学能、电能、原子能替代人类的生物能，人类劳动由体力向脑力转变。而人工智能引领的第三次生产力升级本质上延续了自动化的替代过程，通过计算机、机器人及大数据等信息技术的应用实现对脑力劳动的部分替代。例如半导体材料试验、药品成分配比、电子方案设计等研发工作已经有了人工智能成功应用的案例，提升了新工具和新技术的研发效率，进而实现社会生产效率的跃升。

人工智能挖掘数据资源价值，形成全新的生产要素。互联网时代产生了纷繁的数据信息，而人工智能技术将数据转化为价值。数据资源在人工智能、大数据等技术的价值挖掘下，对个人需求分析、企业商业贸易、国家政策制定都有重要的支撑作用。由于数据资源正处于开发应用的初期，拥有较高的边际效益，对国民经济增长具有较强的促进作用。数据资源的挖掘为人类开辟了认识世界的新维度，为人类带来了新思想与新理论的原料，为科学技术发展带来了新的可能性。

人工智能赋能传统产业，推动国家经济智能化转型。在传统生产层面，人工智能结合物联网技术增强感知能力，拓展智能机器人与智能工厂应用场景，精准追踪与控制生产流程，使得自动化生产与运输成为未来的主流。在供需对接层面，人工智能结合 5G 等宽带网络为生产端和消费端架起了一座桥梁，助力普及个性化、可配置的生产方案，孕育全新的衍生设计方法。按需定制生产有助于缓解低端产能过剩问题，同时为传统产业高端转型构建新的发展路径。在企业经营层

面，智能管理方式重塑传统产业的组织方式，提高员工的生产效率，确保供应链稳健，降低运营成本，优化运作管理流程。

人工智能催生智能产业新业态，激发用户新的消费需求。在创新商业新模式上，智能营销依托大数据分析用户的生活模式，挖掘消费者的潜在需求，帮助企业进一步拓展市场规模。在带动产业新发展上，人工智能有效促进上游关联产业发展，智能芯片、智能存储、智能传感等智能硬件带动基础硬件产业的性能指标升级，智能算法框架、数据交换平台等智能软件营造了开放自由的软件研发环境。在创造消费新空间上，智能音箱等新的人机交互方式以及智能信息服务、智能金融等人性化的互动体验等，激发消费新需求。

（二）人工智能引发经济社会多方面风险挑战

但“每个硬币都有正反面”，人工智能技术在迅速赋能经济社会发展的同时，也引发了安全、社会、法治等不同层面的风险挑战。究其根本原因，在于人工智能所产生的技术异化[1]或将导致治理盲区出现、算法歧视产生、责任主体模糊等多维度治理困境。

一是引发安全风险，技术的不可控将导致侵犯隐私、甚至侵害生命等事件发生。首先，人工智能技术极易被滥用，导致虚假信息频发、不良信息泛滥，加大数字内容治理难度。例如某些应用软件利用深度伪造技术实现图像、音视频的生成或修改（如 AI 换脸），使不良信息内容的“以假乱真”，甚至抹黑政治人物，扰乱网络空间秩序。其次，利用人工智能技术进行数据收集与分析时可能会挖掘出个人敏感信

[1] 《哲学大辞典》对“异化”的释义：异化作为社会现象，与阶级一起产生，是人的物质生产与精神生产及其产品变成异己力量又反过来统治人的一种社会现象。

息并加以利用，在转移与传播阶段可能会在未经用户允许的情况下将数据传递给第三方机构，导致用户隐私泄露。例如某视频类应用在用户录制小视频时窃取其面部等生物识别信息，给用户造成隐私风险；某互联网公司收集儿童和青少年个人信息用于定向推送广告等。再次，智能音箱、自动驾驶汽车等智能产品还未完全成熟，导致安全事故的发生概率大增，甚至威胁到人类生命安全。例如特斯拉、优步等自动驾驶汽车均发生过交通事故；某智能音箱产品曾诱劝主人自杀等。此外，若未来人工智能被武器化、军用化（如侦察机装载 AI 系统），还将导致人类面临生命安全威胁。美国布鲁金斯学会专家于近期指出，若人工智能武器系统被无限制的扩散使用（甚至被极端组织分子使用），将给国际安全局势造成严重混乱与威胁。

二是引发社会风险，就业替代性等导致贫富差距加大、社会不公平性加深。有观点认为，人工智能作为先进生产力之一，将替代技术含量较低的重复性、流程性工作岗位（如电话销售员等），导致结构性失业等问题，甚至有悲观者认为人工智能将会在人类社会中产生一批史无前例的“无用阶层”。例如，牛津大学研究团队利用应用概率分类模型估计了美国 702 种职业未来被计算机替代的可能性，发现其中 47%的岗位处于高度被替代风险；美联社曾调查指出其有近九成的文章可以采用机器写作；据麦肯锡预测，到 2030 年机器人将取代 8 亿人的工作岗位。同时，由于人工智能将影响不同行业企业、不同岗位的分配关系，使社会财富聚集在少数技术领先的行业企业及个人手中，导致收入分配不均衡、数字鸿沟扩大，使技术能力以及受教育程度较

低的人群将在新一轮的社会资源分配中处于劣势地位。

三是引发法治风险，技术自主性等导致责任主体不明、冲击现有法律体系。当前的人工智能技术具备一定程度的自主学习与决策能力，但并不具备法律人格，同时现行法律法规尚未明确界定人工智能的设计、制造、消费、使用等环节的各方主体责任与义务，将导致法律责任认定和划分难度大增。例如，当自动驾驶汽车出现决策错误导致人员伤亡时，当智能医疗助理给出错误的医疗建议导致患者病情加重时，如何界定主体责任等核心问题尚未解决。更进一步，当未来演进至“强人工智能”阶段，若人工智能也能够成为“权责自负”的独立行为主体，届时将会对现行法律法规及其执行机制等造成深远影响。

（三）多重因素提升了人工智能治理难度

相比于数据治理、内容治理等治理主客体明确、治理范围清晰的治理领域，人工智能由于算法不透明、难解释以及技术的跨界传播性和外溢性强等特点，比一般的数字治理范围更广、难度更大。

一方面，算法黑箱性、数据依赖性等导致人工智能决策过程难预测、决策结果难控制。数据、算法是实现人工智能有效赋能的核心要素，但其自身特性大幅增加了技术治理难点。一方面，人工智能算法的决策依赖于海量的数据输入。数据数量、质量、多样性等因素将直接决定算法模型的质量以及所做决策的正确性、公平性、有效性。但同时，数据本质上是社会价值观的缩影与映射，也会包含一些社会偏见。若未能对数据质量进行有效把控，人工智能算法模型便极易习得数据中隐含的偏见谬误，并将其反映到训练结果中，致使人工智能系

统的功能行为变得更加难以控制。另一方面，深度学习、强化学习等人工智能算法模型就像一个“黑箱”，输入数据和输出结果之间存在着人们难以洞悉解释的“隐层”，用户只能被动接受由算法带来的结果而无法洞悉其运行过程。人工智能算法的不透明性，加之其具有自适应、自学习等特性，导致其极易偏离人类预设的目标，其复杂程度愈发超出人类理解和预测范畴。

另一方面，传统治理结构的僵化、治理方法的失效、治理范围的局限等难以满足人工智能治理需求。人工智能的发展创新对技术治理的专业化、智能化等水平都提出了全新要求。因此，传统的技术治理方式方法将很难适用于人工智能治理。**一是传统治理结构不适应**。传统的科层制治理结构难以适应人工智能快速发展而引发的新问题，政府监管治理将难以全面覆盖人工智能所涉及的全部领域，并且自上至下的治理结构将难以准确找到治理对象，无法产生期望的治理效果。**二是传统治理方法难起效**。人工智能具备一定的自主学习及决策能力，其行为结果不能完全归因其背后的程序开发者或者数据提供者。这将导致对责任主体界定的难度增加。**三是传统治理范围有局限性**。人工智能的应用普及有可能会引发结构性失业、数字鸿沟加剧、社会不公平现象增多等全新社会问题。传统治理尚未覆盖以上问题领域，需要构建全新的技术及社会治理体系。

二、全球主要国家和地区高度重视人工智能治理

人工智能正在深刻影响经济社会进程，为全球包容和可持续发展带来全新机遇。但同时，人工智能作为新兴技术所引发的潜在风险，

也需要各国政府及社会各界从人类命运共同体的高度予以回应，推动构建全球人工智能治理体系意义重大。

（一）全球人工智能治理呈现深入发展之势

当前，全球多国已在人工智能伦理原则方面基本达成共识，但人工智能治理还处于初期探索阶段，正在向可信评估、操作指南、政策法规等方面落地实践、逐步深入。

一方面，全球人工智能治理框架体系正在加速构建。美国、欧盟、英国、日本等主要国家及地区已经将伦理治理等纳入人工智能战略中。美国将人工智能治理作为《国家人工智能研究和发展战略计划》八大战略之一；欧盟致力于引领全球人工智能治理，在其发布的《可信人工智能伦理指南》、《欧盟人工智能白皮书》等文件中均强调构建“可信人工智能生态系统”的重要性；英国在其《产业战略：建设适应未来的英国》中，强调要在数据及人工智能安全方面保持领先；日本内阁于 2018 年底发布《以人类为中心的人工智能社会原则》，强调将重视人工智能带来的负面社会影响，积极构建能够使人工智能有效且安全应用的“AI-Ready 社会”。**联合国（ITU）、二十国集团（G20）、经合组织（OECD）等国际组织正积极推动人工智能伦理原则及倡议制定。**如联合国教科文组织于 2019 年 11 月召开第 40 界大会，193 个会员国决定委托该组织就人工智能伦理问题制定第一份全球规范性文件，重点关注人工智能对公平正义和人类权利带来的挑战，并支持推动可持续发展目标方面的国际合作；2019 年 G20 通过《G20 人工智能原则》，倡导以人类为中心、以负责任的态度开发人工智能；

OECD 发布《负责任地管理可信任的 AI 原则》，提出人工智能应遵循的五项伦理原则。**产业界、学术界等积极探索实践人工智能技术治理。**谷歌、微软、百度、旷视、腾讯等科技企业纷纷成立人工智能治理专门机构，通过发布伦理原则、制定内部规范、研发技术工具等举措，形成有效解决相关具体问题的治理体系；牛津大学人类未来研究所发布《AI 治理：研究议题》报告，搭建了“治理愿景、治理技术、治理政策”等三个层次的治理体系；英国阿兰图灵研究所发布《理解人工智能伦理及安全》报告，提出构建负责任 AI 系统的可行性方案；哈佛大学研究团队发布学术论文，设计出模块化的 AI 分层治理模型，并对近期、中期、远期等不同时间维度提出治理目标重点。

另一方面，全球人工智能立法进展相对缓慢，正在从部分领域开始尝试。相较于人工智能伦理规范和指引，全球人工智能立法进展较为缓慢，除在数据隐私保护、算法规制以及重点应用领域立法工作推动相对迅速外，更多的是与现行法律相结合，通过进一步修订或解释进行监管。**数据治理方面**，欧盟《通用数据保护条例》于 2018 年正式实施，随后，美国、日本、巴西、新加坡等国家纷纷出台或修订个人数据保护立法，对人工智能发展起到一定的规范作用；欧委会于 2020 年 11 月通过了《欧洲数据治理条例》建议稿，促进欧盟各成员国之间实现数据共享；2020 年 9 月，我国发布《全球数据安全倡议》，提出有效应对数据安全风险应遵循的三项原则，包括秉持多边主义、兼顾安全发展和坚守公平正义。**算法治理方面**，2019 年，美国参议员提出联邦《算法问责法案（草案）》，建议尽快制定关于“高风险自

动决策系统”的评估规则；2019 年，加拿大出台《自动化决策指令》等对算法审查以规定；欧盟发布《算法责任与透明治理框架》，对算法系统评估流程提出建议。**此外，自动驾驶、金融、医疗、深度伪造等高风险应用领域成为法规制定重点**。自动驾驶方面，2020 年美国发布了《确保美国自动驾驶领先地位：自动驾驶汽车 4.0》，截止目前已有 30 多个州通过自动驾驶相关法规；2017 年德国推出全球首套《自动驾驶伦理准则》，修订《道路交通法》，尝试寻求自动驾驶技术与传统立法的兼容；2018 年，英国出台全球首部为自动驾驶设计保险制度的法律《自动化与电动化汽车法》；2019 年 5 月，日本通过了《道路运输车辆法》的修正案，规定了自动驾驶实用化安全标准；韩国在 2020 年发布《自动驾驶汽车安全标准》，且政府正在为“成为全球第一个将自动驾驶商业化的国家”加紧制定法律监管框架。

（二）不同国家和地区人工智能治理各有侧重

美国高度重视人工智能对民众就业以及国家安全的影响。一方面，美国政府非常重视人工智能对就业带来的影响并提出相应对策。2017 年美国众议院发布《人工智能创新团队法案》，2018 年发布《人工智能就业法案》，提出美国应营造终身学习和技能培训环境，以应对人工智能对就业带来的挑战。另一方面，美国高度重视对国家安全的维护，强调人工智能伦理对军事、情报和国家竞争力的作用，还发布了全球首份军用人工智能伦理原则，如 2018 年，美国设立人工智能国家安全委员会，并承担起考察 AI 技术在军事应用中的风险，考察 AI 技术在国家安全和国防领域的伦理道德问题，以及制定公开训练数据

标准和推动公开训练数据共享等职责；2019 年美国国防创新委员会发布《人工智能原则:国防部人工智能应用伦理的若干建议》,提出“负责、公平、可追踪、可靠、可控”等五大必须遵守的原则。

欧盟将维护人工智能伦理价值观上升至欧洲整体战略层面。欧盟高度重视建立人工智能伦理道德和法律框架，并着力将其推广至整个欧洲，捍卫欧洲价值观，确保人工智能技术朝着有益于个人和社会的方向发展。欧洲科学与新技术伦理小组在《关于人工智能、机器人及“自主”系统的声明》中，提出了一套人工智能发展的基本伦理原则。2018 年，欧盟成立了人工智能高级别专家组，指导相关政策的制定。2019 年 4 月，专家组发布《人工智能伦理准则》，提出建设以人为本的人工智能，列出了可信赖的人工智能系统应满足的 7 个关键要求。同年 6 月，专家组提出人工智能发展的 33 项政策和投资建议。为加强对人工智能的监管力度，欧委会于 2020 年 2 月 19 日发布《欧盟人工智能白皮书》，提出将针对可信赖人工智能建立新的监管框架，建立涵盖事前、事中、事后各个环节的全面监管机制，在确保各种风险和潜在损害最小化的同时，避免过度监管对产业创新发展的阻碍。

英国强调人工智能规范发展并推动 AI 教育及人才培养。英国政府在多份文件和报告中呼吁建立国家层面的人工智能准则与伦理框架。英国下议院在 2016 年发布《机器人技术和人工智能》报告，指出英国应规范机器人技术与人工智能系统的发展。2018 年 1 月发布的《数据宪章》指出，应确保数据以安全和符合伦理的方式使用。2018 年 4 月发布的《英国发展 AI 的计划、意愿和能力》报告提出了关于

AI 准则的五条总体原则，阐明了政府需要考虑的策略性问题。AI 专业人才培养同样受到英国政府的重视。英国提出应加强公民终身再培训，政府应加大技能和培训方面的投资等，在《英国发展 AI 的计划、意愿和能力》《产业战略：建设适应未来的英国》中都提出了具体举措，英国政府还与阿兰图灵研究所合作，向 AI 行业投资用于顶尖人才培养。

法国通过专家研讨、学术辩论等方式深化对 AI 伦理问题的认识。法国国家信息技术和自由委员会在 2017 年举行了涉及 3000 人的 45 次 AI 学术辩论活动，并于同年 12 月推出主题为《人类如何保持优势——算法和人工智能引发的道德问题》的报告，强调了“将人工智能应用于人类服务”的两个基本原则。即“忠诚原则”（意味着算法和 AI 系统应该忠实于用户）和“持续关注和警惕的原则”（意味着参与算法链的所有利益相关者需要处于对可能无法预料的后果的警惕状态）。该报告还提出了六项政策建议，涉及道德教育、提高人工智能系统的可解释性和可审计性，以及促进人类自由等。

德国、韩国等高度重视自动驾驶政策法规制定。早在 2015 年 9 月德国政府便推出《自动化和互联驾驶战略》；2017 年 5 月，德国议会通过了一项由运输部提议的《修订案》，规定在特定时间和条件下，高度或全自动化驾驶系统可接管驾驶人对汽车的控制；2017 年 8 月，德国联邦交通与数字基础设施部推出全球首套《自动驾驶伦理准则》，规定了自动驾驶汽车的行为方式。韩国国土交通部于 2020 年 12 月发布《自动驾驶车辆安全运行准则》，主要包括伦理准则、网络安全准

则、生产与安全准则三大部分。其中伦理准则以确保生命安全为核心，要求自动驾驶车辆在设计和生产时遵循“生命重于财产”、“无法避免事故时应尽量减轻人员伤亡”等原则。

（三）我国人工智能治理稳步推进、成效明显

我国高度重视人工智能规范有序发展，近年来在各方共同努力下，我国人工智能治理稳步推进，发展环境持续优化。

一方面，提出明确的人工智能治理目标。在国务院发布的《新一代人工智能发展规划》中，明确提出人工智能治理“三步走”战略目标：到 2020 年部分领域的人工智能伦理规范和政策法规初步建立；到 2025 年初步建立人工智能法律法规、伦理规范和政策体系，形成安全评估和管控能力；到 2030 年建成更加完善的人工智能法律法规、伦理规范和政策体系。“三步走”战略目标为我国系统构建人工智能治理体系提供了清晰明确的方向指引。工信部发布的《促进新一代人工智能产业发展三年行动计划（2018-2020 年）》中，提出开展人工智能相关政策法规研究，为人工智能产业健康发展营造良好环境。

另一方面，积极探索人工智能治理实践方案。我国政产学研各方积极推动人工智能治理的实践落地。**一是确立治理原则**。我国于 2019 年 6 月发布《新一代人工智能治理原则》，提出八项治理原则以发展负责可信的人工智能；同年 8 月，中国人工智能产业发展联盟发布《人工智能行业自律公约》，旨在引导和规范行业从业者行为；在 2020 年 7 月举办的世界人工智能大会上，成立上海国家新一代人工智能创新发展试验区专家咨询委员会治理工作组，提出构建“1 个平台+4 项工

作+4 个体系”的人工智能治理原则的行动建议。**二是研制技术标准**。2018 年中国国家标准化委员会发布《人工智能标准化白皮书》；目前《新一代国家人工智能标准体系建设指南》正在各部委征求意见；已立项《信息技术人工智能术语》等 4 项国家标准，已发布《面向深度学习的服务器规范》等 11 项团体标准；同时，我国积极推进在 ITU、ISO/IEC 等框架下的国际标准制定工作，如支持国内专家参与 ISO/IEC JTC 1 标准化活动，在 2018 年推动成立 ISO/EC JTC 1/SC42 人工智能分委会，提交了三项国际标准提案。**三是启动相关法规的研究制定**。我国陆续启动数据、算法以及车联网、智能医疗等人工智能相关规范立法工作。如 2020 年 7 月，全国人大常委会第二十次会议审议了《数据安全法（草案）》并公开征求意见；2020 年 2 月，发改委、网信办、工信部等 11 个部委联合印发《智能汽车创新发展战略》，提出到 2025 年要基本形成智能汽车的法规标准、产品监管体系；2020 年经全国人大通过颁布的《民法典》中对保护个人隐私做出明确规定，要求不得泄露、出售或非法向他人提供个人信息等。

三、推动构建人工智能治理体系的思考

人工智能治理是国际组织、国家政府、行业组织、企业、公众等多主体对人工智能研究、开发、生产和应用中出现的公共安全、道德伦理等问题进行协调、处理、监管和规范的过程[2]，是一项复杂的系统性工程，不仅需要提出切实有效的治理举措，还需要根据实际需求及反馈动态调整、持续优化治理路径及策略，实现科技造福人类的治

[2] 中国信息通信研究院《人工智能治理白皮书》

理目标愿景。

（一）秉持科技造福人类、安全与发展兼顾的治理目标

人工智能治理应以“科技造福人类”为目标，既要充分释放人工智能带来的技术红利和价值，也要安全防范、及时应对人工智能可能带来的风险。因此，找准创新发展与有效治理之间的平衡点是关键。对于人工智能发展中出现的问题，既要避免实行过于严苛的限制而抑制创新活力，也不能任其自由泛滥引发更严重的隐患，应给予市场适当的试错、调整空间，同时严守治理原则底线，确保人工智能产业规范有序发展。

（二）打造政府、市场、公众多方协同共治的治理模式

人工智能治理的重要特征之一是治理主体的多元化、治理手段的多样化，其依赖于包括政府、企业等在内的多利益攸关方的参与合作，各司其职、各尽其能，以适当的角色、最佳的手段协同共治。**国家政府及政府间国际组织**在治理中发挥着领导性作用，通过设置专门机构（见附表 1），发布战略政策、制定法律法规等来执行落实治理规则。例如，在 2020 年 G20 数字经济部长会议中，各成员国一致同意愿落实践行《G20 人工智能原则》，共同推动全球人工智能健康发展。**行业组织**是治理的重要推动者，通过制定技术标准、倡导行业自律等方式推动人工智能规范发展。如 ISO/IEC 等标准化组织机构正在制定人工智能技术及产品相关标准；IEEE 还正式发布了《人工智能设计伦理准则》，倡导通过技术手段解决难点问题；日本人工智能学会伦理委员会曾发布《人工智能研究人员应该遵守的伦理指标草案》，旨在

引导和规范研究人员正确处理 AI 引发的伦理道德、安全等问题。**高校及研究机构**致力于推进人工智能治理及伦理研究，为社会提供基础性、公益性服务。例如，哈佛大学曾发起“全球 AI 对话”活动，为社会各界分享 AI 治理观点提供平台；2018 年，欧洲政策研究中心发布的《人工智能的伦理、治理和政策挑战》研究报告，分析研判欧洲目前在人工智能领域的政策和准则现状，提出了欧洲伦理指南的内容框架；2020 年 6 月，清华大学成立人工智能国际治理研究院，面向人工智能国际治理重大理论问题及政策需求展开研究，努力构建公正合理的人工智能国际治理体系。**科技企业及开发者**是践行行业自律自治的中坚力量，通过制定企业伦理原则、研发技术工具等，有力推动治理的落地实践。例如，DeepMind 已宣布成立人工智能伦理与社会部门，旨在补充、配合人工智能研发和应用活动；腾讯在 2019 年发布《智能时代的技术伦理观：重塑数字社会的信任》研究报告，积极倡导“科技向善”的企业责任；旷视于 2020 年 3 月发布《正确使用人工智能的倡议书》，并于同年 6 月与北大联合完成首个 AI 治理商业案例。**公众（包括独立学者）**是治理过程中的重要参与者，拥有监督、意见反馈等权利，应积极参与治理规则制定，适当介入相关监督过程，加强公众对 AI 的理解和信任，更好地保护智能服务消费者的合法权益。2020 年 10 月，我国全国人大法工委在中国人大网公开《中华人民共和国个人信息保护法（草案）》，面向社会公众征求意见建议，保障公众对治理相关立法工作的知情权、监督权、建议权等。

（三）综合运用伦理引导、技术应对、规范立法等治理手段

人工智能治理应建立在对人工智能价值观思考的基础上，通过制定伦理原则、设计技术标准、开发技术工具、确立法律法规等手段举措，有效应对人工智能所带来的风险挑战。**一是伦理原则引导**。对人工智能进行必要的伦理规范约束以使其科技向善、健康发展。相对于立法的反应滞后性，伦理规范约束可以先行预设，能够较为及时地反映出变化的社会关系。国际各方均在积极探索伦理原则相关议题（见附表 2），例如国际网络联盟提出包括“确保系统透明、构建全球管理机制、禁止 AI 军备竞赛”等十项伦理原则；牛津大学、清华大学等高校学者从学术视角讨论 AI 伦理及治理问题。**二是标准规范设计**。人工智能技术及行业标准体系建设工作对人工智能产业发展具有基础性、支撑性、引领性作用，是推动产业创新发展的关键抓手。目前欧、美、日等国高度重视人工智能标准化工作（见附表 3），例如美国于 2016 年发布的《国家人工智能研究与发展策略规划》、欧盟于 2013 年发布的《欧盟人脑计划》以及日本于 2016 年发布的《人工智能/大数据/物联网/网络安全综合项目》均强调围绕 AI 标准规范做更为充足的部署计划。**三是技术工具研发**。智能化数字技术本身也作为一项精准、高效的技术治理工具，正在被用来解决人工智能所带来的风险问题。科技巨头企业正在积极开发运用数据筛选、算法设计、模型优化等技术工具，着力解决诸如隐私泄露、算法偏见、非法内容审核等问题。例如，谷歌已研发出“Explainable AI”工具，帮助用户理解算法的

决策原因、依据和过程，使 AI 算法更加透明、可解释；微软利用“单词嵌入”的自然语言处理工具解决文本搜索中的性别偏见问题；华为利用“差分隐私”“数据过滤”等技术，应用于算法训练等数据使用过程，确保 AI 应用的安全可控。**四是法律法规制定**。人工智能所带来的责任归属界定、隐私泄露等问题不断挑战现行法律法规底线，需要考虑采取包容审慎、灵活弹性的应对方式，既要避免草率立法对人工智能发展带来的阻碍，也要跟上 AI 技术发展节奏，实现“敏捷治理”。目前全球多国已陆续开展对人工智能相关应用场景进行规范立法工作。例如，美国已陆续制定关于自动驾驶、无人机等应用场景的法案；欧盟针对机器人和自主武器系统提出法规倡议；日本则对自动驾驶、著作权、数据利用方面发布相关法案及指南。

（四）构建覆盖全面、反馈及时的治理效果评价机制

针对人工智能技术更迭快、创新发展迅速的特点，治理手段及治理体系机制等也需要对新趋势、新问题进行及时响应、动态优化、持续完善，实现更为敏捷的治理。为此，可适当引入治理效果评价及反馈（或治理绩效考察）等机制，以准确掌握最新治理情况、及时发现问题，以制定更有效的应对策略。例如，对于伦理原则，可评估其可操作性、社会接受度等；对于技术工具，可对其安全性、成熟度、稳健性、有效性等进行评价；对于技术标准规范，可评价其全面性、合理性等；对于法律法规，可评价其社会满意度、是否监督问责等。

四、持续完善我国人工智能治理体系

面向未来，应准确把握新一代人工智能发展特点规律，积极参与

人工智能全球治理议题，借鉴国际社会良好的产业实践和监管经验，加强伦理及治理体系理论研究，探索构建符合中国国情的人工智能治理框架，维护人民利益和国家安全。

（一）加大政策支撑力度，持续优化治理环境

政府应肩负起对构建人工智能治理体系的顶层布局以及对相关技术及服务监督管理的职责，加强资源统筹、部门协作、信息共享，为人工智能营造健康可持续的发展环境。**一是**统筹规划、系统构建人工智能产业规范及保障体系，推动数据治理、算法规制、标准制定、安全评估等能力建设。**二是**推动自动驾驶、智能医疗、智能制造、生物信息识别、服务机器人等重点领域及应用场景的规则制定，加强对新型智能产品或智能服务的风险监测，构建结构合理的人工智能责任划分体系，探索有效可行的监管路径。**三是**加强数据产权问题研究，规范数据权属、使用、交易、共享机制，解决数据所有权、使用权、收益权等问题。**四是**建立和畅通人工智能产品的投诉举报等监督机制，鼓励第三方机构依托标准开展评测，形成标准制定、产业宣贯、测试评估的联动推进策略，开展面向公众的人工智能公共伦理教育。

（二）深化治理规则研究，提升治理能力

进一步提升人工智能治理、数据治理等数字治理能力，持续完善适宜人工智能发展的法律法规环境。**一是**加强人工智能治理研究力量，加大对通用人工智能阶段人机共生社会关系等前瞻性伦理问题研究投入，加强人工智能基础理论、可解释算法、可信任 AI 评估测试、开源开放平台等技术研究，打造具备一定国际影响力的治理专业组织

机构，以务实审慎的态度发展人工智能。**二是**加强行业自律，依托企业、高校、产业联盟等机构，加快完善分享相关操作指引，支持产业界加强实践，推动人工智能伦理及行业自律原则的实施落地。**三是**持续完善人工智能标准体系，鼓励行业组织等制定人工智能产业标准、技术标准和行业规范，推动形成国家标准顶层引导、行业标准先试先行、国际标准积极参与的良好局面。**四是**根据人工智能使用场景、影响范围、可能的危害程度的不同，采用分类治理的思路：对于涉及国家安全、社会稳定等高风险领域（如自动驾驶、智能医疗），加强事前监管与准入限制；对于涉及个人日常消费及服务等风险相对较低的领域（如电子商务、智能家居），采取基于结果的规制思路，侧重事中事后监管。

（三）明确不同阶段治理重点，完善治理路径[3]

应根据人工智能不同的发展阶段，因时因事构建全面、有效的人工智能治理路径，在不同发展阶段明确不同的治理目标和重点。**在近期，应加快制定产品和服务标准，利用“数据治理”推动人工智能治理问题解决**。应重点关注人工智能应用数据对个人权益、国家安全和企业利益带来的影响，以及人工智能在重点行业领域的应用。如借助“数据治理”实现对于算法和应用的治理，加快个人信息保护、数据安全、数据跨境流动、数据共享交换等制度建设；由行业协会或产业联盟出台行业标准规范的方式提升产品和服务的稳定度与质量。**在中远期，应调整责任法律制度，实现法律与伦理原则衔接**。应重新评估自发性

[3] 中国信息通信研究院《人工智能治理白皮书》

的人工智能技术及应用对现行责任体系带来的影响，明确人工智能研发、设计、制造、运营和服务等各环节主体的权利义务，研究在产品责任法律体系中纳入软件等新要素的可行性，同时在部分领域探索算法透明度的要求和程序。

（四）加强国际交流合作，参与全球治理议题

人工智能的技术发展与产业应用是全球高度协作的成果，其未来的发展依赖世界各国优势互补、互利共享、合作共赢。联合国副秘书长法布里齐奥·霍奇尔德曾表示，“没有任何一个单一的国家或公司能够设计出全面、符合所有人预期的 AI 治理体系，我们必须团结起来，为 AI 创建一个可行的国际合作和治理框架”。**一是**秉持合作共赢、共同发展的理念，加强与各国之间的政策沟通，增进理解、建立互信，与全球各方携手构建起共建共享、安全高效、持续发展的人工智能治理环境。**二是**积极参与全球人工智能治理交流对话，推动国际标准化组织讨论和制定全球人工智能伦理规则及标准，鼓励国内外企业及相关机构加强 AI 国际标准化合作，持续提升算法规则、数据使用、安全保障等方面的治理能力。**三是**在人工智能领域建设包容性强的国际治理交流平台，与全球各国共同研究、讨论、采纳各方面好的意见建议，形成国际治理规范共识。**四是**在多边合作机制下，分享治理经验与实践，共同探索合乎人类发展需要的人工智能治理模式。

附表 1. 主要国家及地区设立的人工智能治理专门机构

国家	监管机构名称	监管相关职责	建立时间	负责部门
美国	人工智能专门委员会	负责审查联邦机构的人工智能领域投资和开发方面的优先事项	2018 年 5 月	白宫科技政策办公室、美国国家科学与技术委员会、国防部高级研究计划局等
	联合人工智能中心	监管国防机构人工智能工作	2018 年 6 月	美国国防部
	人工智能国家安全委员会	考察并监督人工智能技术应用在军事中的情况，评估其安全、伦理、对国际法影响等风险	2018 年 11 月	美国众议院武装部队新型威胁与能力小组委员会
欧盟	人工智能高级小组	研究并起草人工智能监管框架、并指导欧洲相关企业进行落实	2018 年 6 月	欧盟委员会
英国	人工智能理事会	监督英国人工智能战略实施并为政府提出建议	2018 年 4 月	英国政府人工智能办公室
	数据伦理和创新中心	审查、监管现有的数据（包括人工智能）治理格局，并就其安全、道德、创新使用为政府提出建议	2018 年 11 月	英国政府
法国	人工智能伦理委员会	监督军用人工智能的发展	2019 年 4 月	法国政府
日本	人工智能技术战略会议	国家层面的人工智能综合管理机构，负责政策及应用的监管	2016 年 4 月	日本政府
印度	人工智能伦理委员会联盟	制定人工智能产品研发标准	2018 年 6 月	印度政府
墨西哥	人工智能办公室	规范人工智能健康发展	2018 年 6 月	墨西哥政府
中国	新一代人工智能发展规划推进办公室	研究人工智能相关法律、伦理、标准、社会问题以及治理议题	2017 年 11 月	国家科技体制改革和创新体系建设领导小组

来源：公开资料整理

附表 2. 人工智能伦理原则及规范相关文件举例

文件名称	核心内容	发布时间	发布机构	机构所在国家
《阿西洛马人工智能原则》	确保人工智能系统安全、审判透明、与人类价值观一致、保护个人自由和隐私、可控、避免军备竞赛、造福人类	2017 年 1 月	未来生命研究所	美国
《人工智能设计伦理准则》	保护人权、保障民生福祉、可控可责、确保运行透明、慎用	2017 年 12 月（第二版）	电气和电子工程师协会	
《人工智能使用原则》	社会有益、避免偏见、人类担责、保护隐私、保证安全	2018 年 6 月	谷歌	
《人工智能伦理原则》	负责任；公平；可追溯；可靠；可管理；呼吁国防部增加对 AI 研究、培训、道德和评估的投入	2019 年 10 月	美国国防部	
《可信任人工智能伦理指南》	遵守法律和伦理道德、尊重人类自由自治、受人监管、避免伤害、保证公平、稳定可靠、保护隐私、透明可释、可审核评估、可问责、确保社会福祉	2019 年 4 月	人工智能高级别专家组	欧盟
《机器人和机器系统的伦理设计和应用指南》	机器人不能伤害人类、人类要为机器人负责、避免性别和种族歧视	2016 年 4 月	英国标准协会	英国
《英国发展人工智能的计划、意愿和能力》	确保人类共同利益、保证公平、容易理解、保护隐私、普及教育、避免伤害欺骗人类	2018 年 4 月	英国上议院	
《日本人工智能学会伦理准则》	贡献人类、遵守法律、尊重隐私、公平公正、保证安全、肩负社会责任	2017 年 2 月	日本人工智能伦理委员会	日本
《自动驾驶伦理准则》	保证交通参与者安全、驾驶系统需要官方批准监管、禁止将人群属性作为评价标准、禁止量化生命价值、责任共担	2017 年 8 月	德国交通与数字基础设施部	德国
《法国人工智能发展计划》	算法透明、责任承担、成立人工智能伦理委员会、组织伦理公共辩论	2018 年 3 月	法国政府	法国
《可靠的人工智能草案蒙特利尔宣言》	保证民生福祉、公平正义、保护隐私、促进民主、共担责任	2018 年 12 月	蒙特利尔大学	加拿大

文件名称	核心内容	发布时间	发布机构	机构所在国家
《人工智能伦理十大原则》	系统透明、使用“道德黑匣子”、服务于人和地球、受人控制、无偏见、福泽全人类、保证公平和自由、建立全球管理机制、遵守法律、禁止军备竞赛	2017 年 12 月	国际网络联盟	国际组织
《负责任地管理可信任的 AI 原则》	包容性增长、可持续发展和福祉、以人为本的价值和公平、透明可解释、安全稳健可靠、负责任	2019 年 5 月	经济合作与发展组织	
《人工智能北京共识》	有益社会、增进生态福祉、服从人类利益、设计合乎伦理、包容多样、风险管控、避免滥用误用、加强教育培训、构建全球治理框架	2019 年 5 月	北京智源人工智能研究院	中国
《面向儿童的人工智能北京共识》	社会各界应高度重视人工智能对儿童的影响，发展对下一代负责任的人工智能。人工智能的发展应保护和促进儿童的权益，避免剥夺和损害儿童的权利，助力实现儿童健康成长	2020 年 9 月		
《发展负责任的人工智能：我国新一代人工智能治理原则》	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	2019 年 6 月	国家新一代人工智能治理专业委员会	
《关于加强科技伦理建设践行科技向善理念的建议》	科技向善、负责任创新、创新与伦理并重、加强科技伦理的制度化建设、加强科技伦理的教育宣传	2019 年 3 月	腾讯	
《人工智能应用准则》	从正当性、人的监督、技术可靠性和安全性、公平和多样性、问责和及时修正、数据安全与隐私保护六个维度，对人工智能正确有序发展作出明确规范	2019 年 7 月	旷视	
《协同落实人工智能治理原则的行动建议》	提出了“一平台、四工作、四体系”的系统落实人工智能治理原则的行动方案建议	2020 年 7 月	上海国家新一代人工智能创新发展试验区专家咨询委员会治理工作组	

来源：公开资料整理

附表 3. 人工智能标准化文件举例

文件名称	核心内容	发布机构
《ISO/IEC 20005》（2013 年）	传感器网络标准化：术语与词汇、智能传感网络协同信息处理支持服务和接口	国际标准化组织、国际电工委员会（ISO/IEC）
《ISO/IEC 30122》（2016 年）	人机交互标准化：框架与通用指南、构建与测试、翻译与本地化、语音命令注册管理	
《ISO/IEC 19944》（2017 年）	云计算标准化：互操作与可移植、跨设备数据与云服务的数据流	
《算法透明和可责性声明》（2017 年）	充分认识算法歧视、明确数据来源、提高可解释性与可审查性、建立严格的验证测试机制	美国计算机协会（ACM）
《IEEE P7000》（2016 年）	系统设计期间解决伦理问题的模型过程的标准	电气和电子工程师协会（IEEE）
《IEEE P7001》（2016 年）	自主系统的透明度的标准	
《IEEE P7002》（2016 年）	数据隐私处理的标准	
《IEEE P7003》（2017 年）	算法偏差注意事项	
《IEEE P7006》（2017 年）	个人数据人工智能代理标准	
《中文语音识别系统通用技术规范》《中文语音合成系统通用技术规范》《自动声纹识别（说话人识别）技术规范》《中文语音识别互联网服务接口规范》《中文语音合成互联网服务接口规范》	语音交互系列标准	全国信息技术标准化技术委员会
《共享学习系统技术要求》（2020 年）	共享学习的技术框架及流程、技术特性、安全要求	中国人工智能产业发展联盟

来源：公开资料整理

编写组名单

指　　导：

辛勇飞　中国信通院政策与经济研究所所长

王爱华　中国信通院副总工程师

何　伟　中国信通院政策与经济研究所副所长

何　霞　中国信通院政策与经济研究所副总工程师

策划主编：

刘铁志　中国信通院政策与经济研究所战略部主任

胡昌军　中国信通院政策与经济研究所战略部副主任

报告负责：

韩凯峰　中国信通院政策与经济研究所战略部研究员、统稿人

詹远志　中国信通院政策与经济研究所战略部研究员

成　　员：

张芳纯　中国信通院政策与经济研究所战略部研究员

付　江　中国信通院政策与经济研究所战略部研究员

王亦菲　中国信通院政策与经济研究所战略部研究员

刘媛媛　中国信通院政策与经济研究所战略部研究员

中国信息通信研究院 政策与经济研究所

地址：北京市海淀区花园北路 52 号

邮政编码：100191

邮件：hankaifeng@caict.ac.cn

联系电话：010-62301901

传真：010-62302476

网址：www.caict.ac.cn

人工智能与经济社会研究中心

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010-62301901