CAICT 中国信通院

集智 蓝皮书 No.202415

人工智能治理蓝皮书

（2024 年）

中国信息通信研究院

2024年12月

版权声明

更名声明

原“集智”白皮书更名为“集智”蓝皮书。“集智”蓝皮书将继续秉承原有的编撰理念和高质量标准，致力于提供有价值的信息和洞见。

前　言

人工智能治理是各方为解决人工智能领域风险分担和利益分配问题，通过政策制定、法律监管、伦理指导等手段，对人工智能的研发、应用等行为进行全面管理和调控的过程。人工智能治理应为人工智能领域发展和安全问题建立有效的风险矫正机制、利益分配机制及机构协调机制，积极应对发展不平衡、规则不健全、秩序不合理问题并促进全球协商合作。

近一年来，全球人工智能治理面临复杂多变新形势。全球人工智能行业实现高速增长，在产业规模、投融资、企业数量等方面表现明显，具身智能、数字人等人工智能应用将驱动下一轮产业洗牌。通用人工智能技术飞速跃迁敲响安全警钟，幻觉难消除、场景难限定、责任难追溯等特性放大了虚假信息、隐私侵犯、网络犯罪等现实风险。同时，全新的人机交互模式开启了“人类外脑”时代，在未来发展中可能引发情感依赖、劳动替代、生存性风险等方面的人机伦理风险，不断挑战政府传统监管模式、加剧国际合作协调难度、考验供应链主体自治能力。

本报告从新技术革命引发的经济社会发展变革和历史经验出发，在“以人为本、智能向善”理念指引下，基于供应链条、价值链条、全生命周期链条等底层逻辑，从 What、Why、Who、How 四个维度搭建了人工智能治理体系框架。在梳理人工智能概念分歧、人工智能风险谱系基础上，框架从安全和发展两个维度提出四组议题。安全侧来看，安全可控是人工智能治理的底线基础，强调基于风险的全生命

周期治理思路，亟需对透明度、红队测试、评估评测等工具细化落地；伦理先行是人工智能治理的价值导向，需提前研判人工智能对人类生活、生产乃至生存问题带来的冲击，强调敏捷治理、动态监测。发展侧来看，负责任创新是人工智能治理的源头根本，存在算力供给不平衡、高质量数据集建设、开源模型生态治理等问题，提出基于产业链的“要素+场景”治理；可持续发展是人工智能治理的终极要求，提出公平普惠的包容性治理，围绕经济、社会和环境三个维度，协力应对数字鸿沟、能源短缺等问题，推动联合国2030年可持续发展目标实现。

从落地实践来看，主要经济体人工智能治理模式初具雏形，在监管模式、治理重点、实践策略等方面日益成熟完善。企业主体、专业机构、产业联盟等主体积极实践，在安全技术研发、资源共享等方面发挥关键作用。大国间围绕人工智能安全合作共识初建，各方在议程设置、规则塑造等多方面进程提速，联合国系统积极酝酿新的人工智能治理协调机制，各双多边机制推动高层次承诺向可执行政策落地。

展望未来，人类社会将迈向更深层次的智能化发展阶段，需要系统谋划、综合施策，在伦理监测、制度设计、监管模式、国际合作等方面持续改革创新。中国信息通信研究院持续跟踪评估人工智能治理焦点议题和趋势进展，尝试提出人工智能治理的体系框架，期待为推进各方交流讨论、促进合作共赢贡献绵薄之力。

目 录

图 目 录

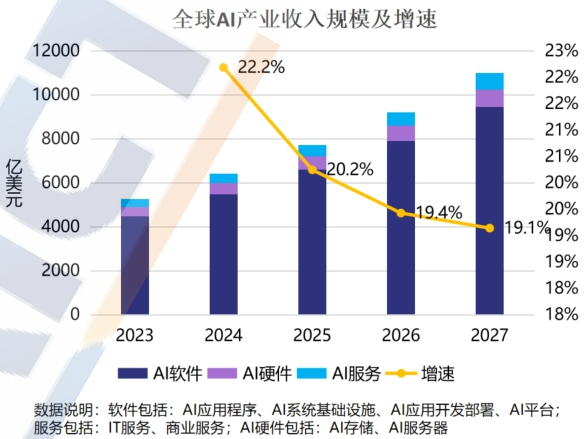
表 目 录

一、人工智能治理面临复杂多变新形势

人工智能技术颠覆性、跨越式突破引发通用人工智能新一轮发展热潮，推动全球人工智能治理进入规则构建的关键阶段。人工智能领域发展不平衡、规则不健全、秩序不合理等问题日益显现，全球人工智能治理面临前所未有的挑战。

（一）人工智能行业高速增长，商业生态分化形态初现

大模型等新技术带动全球人工智能行业实现高速增长。从产业规模来看，受人工智能存储、服务器等基础设施市场拉动，2024 年全球人工智能产业收入高速增长。据 IDC 预测，2024 年全球人工智能产业收入规模达 6421.8 亿美元，同比增长 22.2%。[1]**从企业财务表现看，**人工智能技术研发和应用落地有效推动了巨头企业营收增加，微软、谷歌、亚马逊营收增速均超 10%，AI 云服务、AI 助手及 AI 广告投放在其中发挥关键作用。

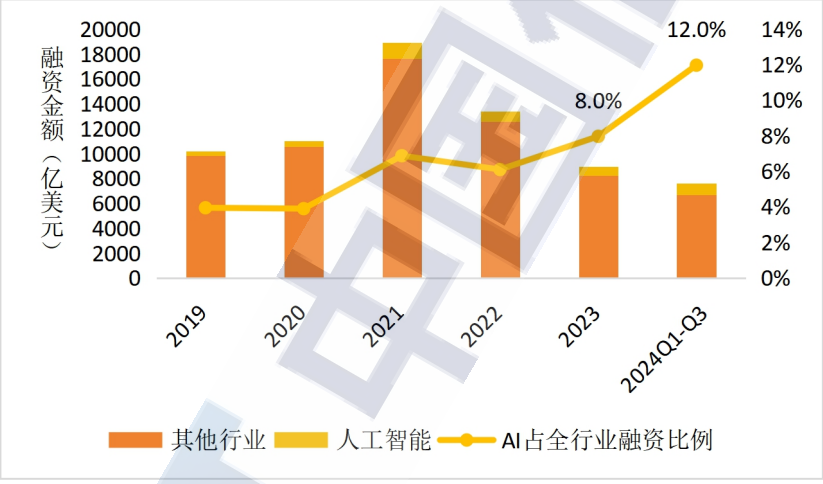


来源：根据 IDC 数据整理

图 1 全球人工智能产业规模-亿美元

[1] IDC Asia/Pacific AI Maturity Study 2024. https://www.intel.in/content/dam/www/central-libraries/us/en/documents/2024-04/idc-infobrief-asia-pacific-ai-maturity-study-2024-india.pdf

从投融资来看，2022-2024 年前三季度，全球投资持续低迷，融资金额持续下降，相比较而言，**人工智能领域融资复苏，**全球 AI 融资占全行业融资比例持续上升，从 2022 年的 6.1%上升至 2024 年前三季度的 12.0%。**人工智能领域巨额融资金额大幅上升。**2024 年前三季度，全球巨额融资金额为 265.6 亿美元，同比上升 148%，融资笔数为 73 笔，同比上升 40%。**大模型领域融资金额大幅上升。**2024 年前三季度全球大模型企业风险融资额为 162 亿美元，已约为 2023 年的两倍，投资笔数达 165 笔。

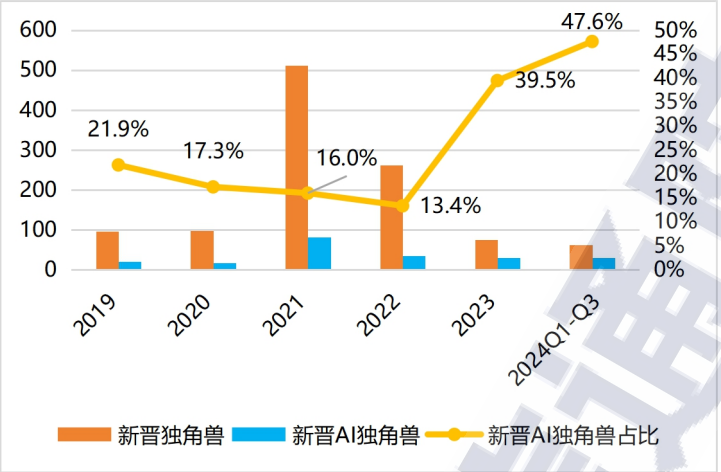


来源：根据 CBInsights 数据整理

图 2 2019-2024 Q3 AI 领域全球融资占全行业融资比例

从企业数量来看，截至 2024 年第三季度，全球人工智能企业数量为 31206 家，其中美国企业 10840 家，占全球总数的 35%，中国企业 4676 家，占全球总数的 15%。AI 独角兽数量大幅增加，截至 2024 年第三季度，全球 AI 独角兽共 255 家。其中，2024 年 AI 独角兽新增数量达 30 家，占所有新增独角兽企业数量的 47.6%。AI 独角兽的

主要赋能领域为商业智能、AIGC 及大模型、医疗等板块。[2]



来源：中国信息通信研究院

图 3 新晋 AI 独角兽数量及占比

人工智能技术创新需求持续加大，产业链条拉长、商业生态分化等特征显现。自 2022 年 11 月 ChatGPT 发布以来，人工智能已被应用于金融、法律、设计等多个行业。人工智能产业链条从大多数情况下由单一企业完成，发展为在开发、调优、应用等不同环节，由不同企业合作完成。围绕生态主企业构建的商业生态开始出现分化，开源与闭源、通用与专用等不同路线均呈现竞争与融合并存态势。

（二）人工智能技术深刻变革，挑战政府传统监管能力

以大模型为代表的新一轮人工智能具有强生成、强推理、强交互等能力特征，技术深度变革为人工智能政府监管带来新挑战。一是体现为新内容生产方式，从“人和信息”的链接转向“人和答案”的链接，规模化提升生成效率，内容生成机制成为人类难以破解的黑箱。**二是**变革人机交互模式，从传统的复杂繁琐、多样化的交互接口转变

[2] AI 100: The most promising artificial intelligence startups of 2024 - CB insights research. https://www.cbinsights.com/research/report/artificial-intelligence-top-startups-2024/

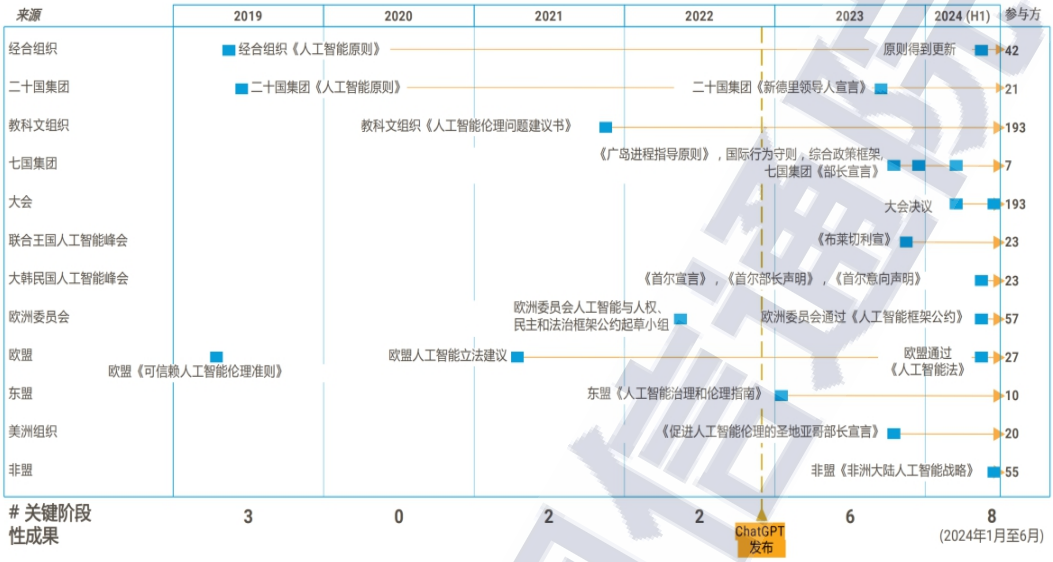
为统一便捷的自然语言接口。人工智能脱离工具角色，成为能自主完成工作流程的“人类外脑”。**三是**通用泛化水平急剧攀升，未来可能在产业发展中发挥基础设施和数字技术市场“看门人”作用，基础性、公共性、外溢性将显著增强。在此背景下，**传统政府监管模式和工具面临重大挑战。一是**事前风险审查缺乏统一标准，政府亟待优化评估审查工具。斯坦福大学《2024 年人工智能指数报告》[3]指出，人工智能严重缺乏标准化评估，OpenAI、谷歌和 Anthropic 等企业根据不同的负责任人工智能基准测试他们的模型，导致难以对头部人工智能模型的风险和局限性进行比较。**二是**事中介入机制缺乏实践经验，政府亟需提升常态化监测水平。各主要国家未就事中监测和介入手段积累有效经验，对人工智能的事中监管尚处初期阶段。**三是**事后调查溯源取证困难，政府尚缺乏有力的核查核验工具。监管部门难以核查企业真实合规情况，在生成式人工智能版权纠纷、内容安全等问题上取证困难，尚缺乏成熟完备的技术溯源措施。

（三）国际合作取得关键成果，各国治理共识差异并存

近一年来，主要经济体在联合国、二十国集团、经合组织、国际电联、金砖国家等多双边机制中围绕全球人工智能治理开展对话协商，取得了部分关键成果，如联合国《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》《加强人工智能能力建设国际合作决议》、七国集团《广岛进程指导原则》等。然而，全球各方因政治制度、法律体系和文化传统的不同，对人工智能治理模式和需

[3] AI index report. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/research/ai-index-report

求有着不同认识，人工智能安全和发展张力加大，进一步增加了共识凝聚的难度，削弱了国际协调合作的动力。



来源：HLAB-AI《治理人工智能，助力造福人类》终期报告

图 4 2019-2024 年 6 月区域间和区域性人工智能治理举措及关键成果

在治理需求方面，人工智能安全可控和发展优先路线并存。一方面，国际社会普遍关注人工智能风险问题，将“安全、可靠和值得信赖”理念纳入多部共识文件；同时探索通过英国、韩国、法国系列人工智能峰会，推进全球安全研究网络和评估评测合作。**另一方面，**各国对技术发展不平衡带来的“智能鸿沟”问题表示忧虑。例如 G20 主席国巴西将“人工智能资产和基础设施分布不均”作为优先议题。[4]东盟发布《东盟人工智能治理与伦理指南》，阐述东盟“发展优先”的治理观，提出支持人工智能人才培养、人工智能创新等国家级建议。**在治理机制方面，多边机制和多利益攸关方均发挥重要作用。**部分国家认为政府对人工智能等新技术知识认识不足，应推动发挥企业、第

[4] 《2024 年二十国集团数字经济部长会议在巴西马塞约召开》，载人民邮电报，https://mp.weixin.qq.com/s/1mZkQCL4Z8DfFEJs9eUXpg.

三方机构等多利益攸关方作用。发展中国家则普遍面临基础能力不足等挑战，希望通过多边机制参与全球人工智能治理。人工智能治理评估指数（AGILE 指数）报告[5]显示，发展中国家在技术、资金和人才方面的不足，限制了它们在人工智能国际治理中的参与度和影响力。总体来看，世界经济、技术不稳定风险仍存，各方需高度认识人工智能技术带来的巨大挑战。

（四）人工智能技术风险加剧，亟需搭建完善治理框架

1. 人工智能风险的系统认识

人工智能风险可划分为内生风险和衍生风险。**内生风险主要指人工智能系统自身所存在的风险，包括数据、框架、模型等自身安全带来的风险问题。**比如模型存在缺陷、后门被攻击利用的风险，具体指人工智能算法模型设计、训练和验证的标准接口、特性库和工具包，以及开发界面和执行平台可能存在逻辑缺陷、漏洞等脆弱点，还可能被恶意植入后门，存在被触发和攻击利用的风险。再比如人工智能的生成“幻觉（Hallucination）”问题导致生成内容不可信风险。生成幻觉通常指模型按照流畅正确的语法规则产生的包含虚假信息甚至无意义内容的文本。OpenAI 公司前首席技术官 Mira Murati 指出，底层大型语言模型的最大挑战是会编造错误的或不存在的事实。[6]

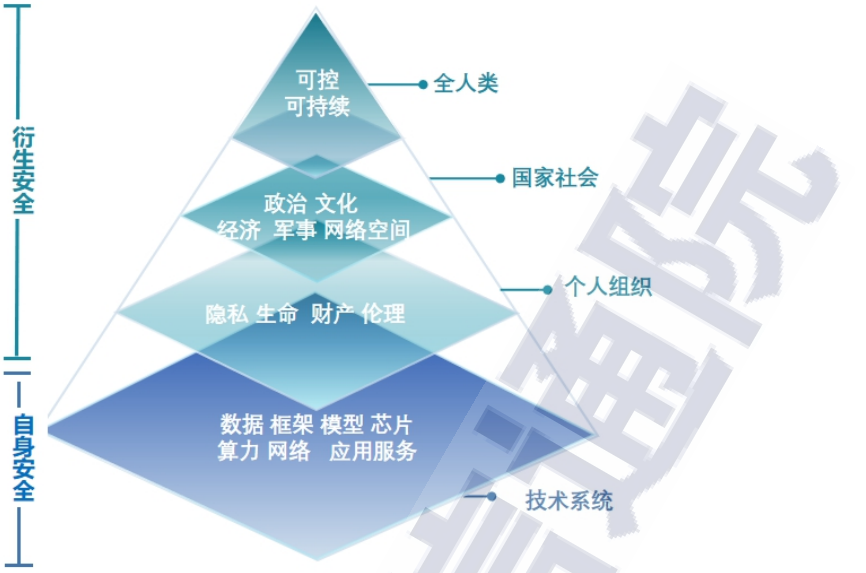
衍生风险包括个人、社会、国家和全人类等多个层面。个人层面

[5] 《全球人工智能治理评估指数(AGILE 指数)正式发布，首次评估解码全球人工智能治理新格局》，载深圳市人工智能产业协会官网，https://www.szaicx.com/page142?article_id=14220.

[6] Radulovski, A. (2024, January 4). Unveiling 7 interesting facts about Mira Murati, OpenAI’s chatgpt creator and Trailblazer in Tech. Women in Tech Network. https://www.womentech.net/en-us/blog/unveiling-7-interesting-facts-about-mira-murati-openais-chatgpt-creator-and-trailblazer-in.

包括对个人隐私、自主性等人格尊严带来的挑战。2024 年 10 月，美国美国佛罗里达州奥兰多地方法院受理“全球首例 AI 机器人致死案”，14 岁少年的母亲梅根·加西亚指控 AI 公司 Character.ai 存在管理疏忽，导致其聊天机器人产品诱导青少年开枪自杀。**社会国家层面包括对就业、公共安全、国家安全等带来的风险。**例如人工智能冲击就业市场，加剧财富分化与社会不公。据高盛研究报告分析，以美国为例，46%的行政工作和 44%的法律工作将受到较高程度的影响。[7]人工智能可以被用于制作虚假文本、音频、视频等深度伪造内容，进而实施政治干预、煽动暴力和犯罪等破坏公共利益、侵害国家安全行为。**全人类层面包括能源消耗、数字鸿沟、生存性风险等问题。**例如，模型训练导致大量资源消耗，抬高碳排放水平。研究人员指出，GPT-3 模型能耗相当于 1287 兆瓦时的电力，同时产生 552 吨二氧化碳。[8]近一年来，前沿人工智能引发的生存性风险受到关注，从英国主导的系列 AI 安全国际峰会以及 OpenAI、Anthropic、DeepMind 等企业认识来看，**化学、生物、放射和核（统称 CBRN）滥用、自主复制等被认为是典型的生存性风险。**Anthropic 与顶级生物安全专家一起对其模型进行红队测试后，警告生物滥用风险可能在未来 2–3 年内成为现实。

[7] Artificial Intelligence. Goldman Sachs. https://www.goldmansachs.com/insights/artificial-intelligence.
[8] Stanford HAI. (2023). Artificial Intelligence Index Report 2023. p120. Retrieved from https://aiindex.stanford.edu/report/.



来源：中国信息通信研究院

图 5 人工智能风险视图

2. 人工智能治理的体系框架

人工智能治理攸关全人类命运，具有典型的跨国界、跨学科、分散性、动态性等特征。联合国《全球数字契约》指出，人工智能具有积极和消极两方面颠覆性潜力，需要全球性和多学科的对话，以审视、评估和调整人工智能应用。近年来，全球部分国际组织、政府、企业、研究机构等均尝试从不同视角维度搭建人工智能治理体系框架。

表 1 本报告所参考的人工智能治理框架素材

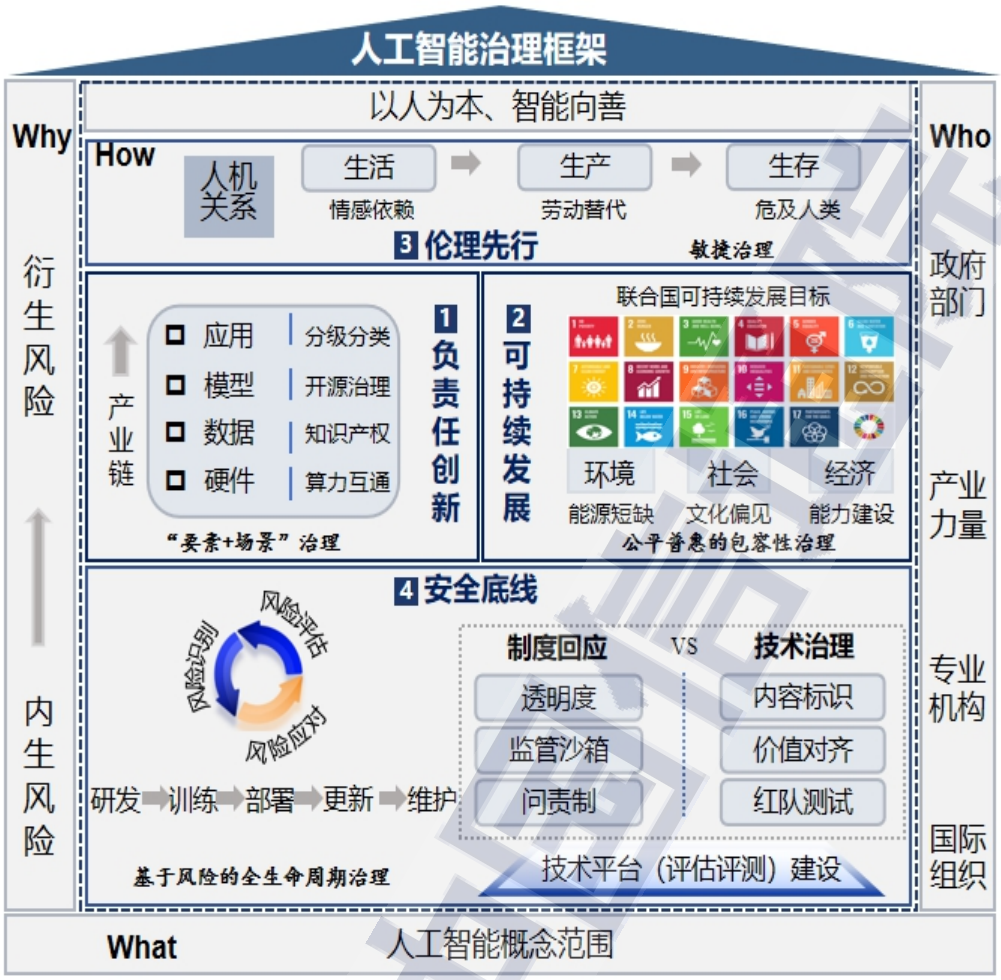
序号	框架名称	发布机构	发布时间	类型
1	联合国系统人工智能治理白皮书：分析联合国系统的制度模式、功能和现有制度适用于人工智能治理的国际规范框架	联合国人工智能机构间工作组	2024 年 5 月	国际组织
2	G7 广岛进程守则框架	G7	2023 年 5 月	国际组织
3	人工智能系统的分类框架	OECD	2022 年 3 月	国际组织
4	东盟关于人工智能治理与伦理指引文件	东盟	2022 年 11 月	国际组织

5	实施人工智能治理：从框架到实践	欧委会	2022 年 12 月	区域性组织
6	人工智能风险管理框架 1.0	美国 NIST	2023 年 1 月	科研院所
7	生成式人工智能模型人工智能治理框架：促进可信赖的生态系统	新加坡 AI Verify 基金会与新加坡信息、通信及媒体发展管理局联合发布	2024 年 5 月	非营利组织/政府机构
8	可信人工智能总体框架	中国信息通信研究院	2021 年 7 月	科研院所
9	人工智能风险管理体系	中国信息通信研究院	2023 年 12 月	科研院所
10	面向产业的人工智能风险治理实践框架	中国信息通信研究院	2024 年 12 月	科研院所
11	全周期人工智能治理框架	Presidio	2024 年 1 月	企业
12	全生命周期风险治理框架	阿里巴巴	2023 年 8 月	企业
13	前沿安全框架（1.0 版本）	DeepMind	2023 年 7 月	企业
14	负责任的扩展政策	Anthropic	2023 年 4 月	企业
15	人工智能治理框架与实施路径	清华大学	2024 年 7 月	学术机构
16	人工智能治理框架	图尔库大学人工智能治理和审计组织（AIGA）	2022 年 6 月	学术机构
17	人工智能治理的分层模型	哈佛大学	2021 年 9 月	学术机构
18	卡内基人工智能国际治理框架	卡内基	2020 年 10 月	非营利组织

来源：中国信息通信研究院

从具体特点来看，现有治理框架分为以下几种类型：**一是全面宏观的治理框架，对于推进国际合作具有重要意义，**例如《联合国系统人工智能治理白皮书》所包含的风险类型和议题最为广泛，既有传统

的透明度、数据和网络安全风险，也有近期广受关注的劳动替代、环境伦理等风险；既有监管执法难等挑战，又有发展方面的包容性、数字鸿沟等问题。**二是注重基于生产流程的全生命周期治理框架，体现人工智能治理的流程性。**例如美国 NIST 框架中的全生命周期包括设计、收集和处理数据、建立、使用和修改模型、开发和使用、操作和监控、使用或被影响的群体和生态等环节。中国信息通信研究院“可信人工智能总体框架”、“人工智能风险管理体系”均涵盖规划设计、数据处理、模型建设、测试验证、部署上线等阶段。**三是提出基于产业链或社会影响的分层治理，体现人工智能治理的产业逻辑。**例如《联合国系统人工智能治理白皮书》提出基于计算机硬件-云平台-数据和人工智能模型-应用程序-服务的人工智能价值链，并针对不同环节提出存在的风险问题。哈佛大学 AI 治理框架提出社会和法律层、道德层和技术层三层框架等。芬兰图尔库大学 AI 治理框架提出人工智能系统层、组织层和环境层三层框架。**四是侧重于某一类人工智能的风险管控框架**，如遵循韩国首尔 AI 安全峰会承诺，OpenAI、Anthropic、Deepmind 等企业均提出前沿人工智能的风险管理框架。以 Deepmind 的《前沿人工智能安全框架》为例，提出风险预警评估、风险缓解措施、明确风险阈值，以及问责制、透明度等保障机制。



来源：中国信息通信研究院

图 6 人工智能治理框架图

在梳理上述治理框架基础上，本报告从新技术革命引发经济社会发展变革和历史经验出发，在“以人为本、智能向善”理念指引下，基于产业链条、治理链条、价值链条等底层逻辑，从 What、Why、Who、How 四个维度搭建了“1244”治理框架，即一组概念、两类风险、四类主体、四组议题，力求框架的全面性、系统性和前瞻性，希望为加深人工智能治理体系研究、推进各方人工智能治理共识提供框架基础。**从治理理念来看，**2023 年 10 月，习近平总书记提出《全球人工智能治理倡议》，强调坚持以人为本、智能向善的理念和宗旨，为人

工智能治理指明了根本性方向。框架整体以“以人为本、智能向善”为基本理念，坚持统筹安全和发展、活力和秩序、效率和公平的关系。**从治理主体来看，**倡导多元主体敏捷互动治理、合作治理的思路，不仅改革创新政府横向监管模式，也为企业、专业机构等参与纵向治理提供充足空间，推进多元主体平等交流和协商对话。

从治理议题来看，在发展方面，负责任创新是人工智能治理的源头根本，存在算力供给不平衡、高质量数据集建设、开源模型生态治理等问题，提出基于产业链的“要素+场景”治理；可持续发展是人工智能治理的终极要求，提出公平普惠的包容性治理，围绕经济、社会和环境三个维度，协力应对数字鸿沟、能源短缺等问题，推动联合国 2030 年可持续发展目标实现。**在安全方面，**安全可控是人工智能治理的底线基础，强调基于风险的全生命周期治理思路，亟需对透明度、红队测试、评估评测等工具细化落地；伦理先行是人工智能治理的价值导向，需关注通用人工智能等技术发展带来的人类主体性危机，强调通过敏捷治理、动态监测的思路，提前研判人工智能对人类生活、生产乃至生存问题带来的冲击。

二、人工智能治理的核心议题进展及趋势

（一）基于产业链的治理：全要素推进负责任创新

负责任创新是人工智能治理的源头根本，涉及芯片、数据、模型、应用等产业链各个环节。当前，国际社会共同关心高质量数据集供给等问题，为人工智能创新发展谋求要素支撑和制度空间。

1. 数据层，高质量数据集的供给存难题

高质量数据集直接影响人工智能的最终性能。**数据集来源合法性是人工智能研发面临的首要障碍。**OpenAI 在《Scaling Laws for Neural Language Models》中提出大语言模型所遵循的“缩放法则”[9]，即增加训练数据量有助于提升预训练模型的效果。实践中，**由于严苛或模糊的制度要求，企业往往难以通过正常、合法的途径获取数据，面临较大个人信息和知识产权侵权风险。在个人信息保护方面，**知情同意授权规则难以得到实施，同时难以援引订立合同所必须、履行法定职责等其他合法性基础。而匿名化处理的具体要求和标准尚不明确，导致开发企业难以把握采集、利用用户数据的合法边界。**在知识产权规则方面，相关诉讼纠纷和监管处罚不断。**美国各地法院已受理超 19 起未经授权使用版权材料进行模型训练的诉讼。法国竞争管理局因谷歌未向训练数据版权人支付费用，对其处以 2.5 亿欧元罚款。[10]YouTube 首席执行官尼尔·莫汉（Neal Mohan）表示，如果 Sora 文生视频大模型被发现使用了 YouTube 来源数据，将违反该平台的服务条款。我国国内也出现“Trik 模型侵权案”“奥特曼侵权案”[11]等相关案件。

多国在制度层面放宽数据来源限制，解决高质量数据集关键堵点问题。欧盟《人工智能法》序言第 140 条规定，参与沙盒测试的企业可以将基于其他合法目的收集的个人数据用于人工智能训练，而不需要用户同意。2024 年 7 月，韩国发布了《关于处理公开数据以开发和服务 AI 的指南》，允许基于“合法利益”使用公开数据进行 AI

[9] Scaling laws for neural language models. OpenAI.https://openai.com/index/scaling-laws-for-neural-language-models/.
[10] 参见《法国监管方对谷歌侵犯版权行为罚款 2.5 亿欧元》，https://m.gmw.cn/2024－03/22/content_1303692195.htm.
[11] 广东省汕头市中级人民法院民事判决书，（2023)粤 05 民终 945 号。

训练和服务开发。**知识产权方面，韩国、德国和英国等国家持开放态度。**2024 年 9 月，德国汉堡地方法院在 LAION 案一审判决中首次确认，AI 模型训练可适用文本与数据挖掘（TDM）版权保护的例外[12]。2024 年 10 月，英国政府就 AI 训练中的“选择退出”版权制度进行公开咨询，该制度以合理使用为原则，权利保留为例外。**美国司法判例对使用版权作品的态度较为宽松。**2024 年 11 月，纽约联邦法院驳回了 Raw Story 和 Alternet 针对 OpenAI 聊天机器人训练数据提起的版权诉讼。[13]法官认为，原告寻求救济的真正原因是 OpenAI 未经授权使用其文章进行 AI 训练且未提供补偿，但这种损害缺乏版权法依据，因此原告不具备提起诉讼的资格。

2. 模型层，开源模型的生态治理引争议

开源模型具有促进人工智能创新的公共属性，可显著降低创新成本，加速人工智能迭代速度。例如，Stable Difiusion 模型通过开源吸纳全球创新力量，在不到两个月的时间内，实现了图片生成速度 15%以上的提升。而在实践中，**大模型开源生态受多方因素裹挟限制。一方面，模型开源实际受商业利益影响，未完全遵守透明和协作原则。**如 Meta、Amazon 等企业将开源作为追赶行业领先者的手段，通过开源部分模型或部分模块，促进行业内形成利已生态，同时部分关键技术因商业秘密、隐私保护等原因保留为闭源，从而增强自身竞争优势。**另一方面，开源模型的资源支持和技术积累难度更大。**大模型对数据集、算力和高效分布架构等基础资源需求较高，头部科技公司主导的

[12] Hamburg Regional Court, Germany [2024]: Robert Kneschke v. LAION e.V., Case No. 310 O 227/23
[13] https://roninlegal.in/raw-story-media-amp-alternet-media-inc-v-openai-legal-implications-for-ai-and-copyright/

闭源模型更容易汇聚丰富的资源用于持续性技术迭代，并能迅速响应市场需求变化。

前沿人工智能开源可能因滥用和失控引发灾难风险。美国生命未来研究所认为，开发者在开源模型后，会丧失对模型的控制和限制能力，开源模型可能会被滥用于恶意目的，并造成大规模的意外伤害。[14]GovAI 发布报告指出，高能力模型可能被用于生产生化武器或进行网络攻击，开源可能加剧极端风险发生。**对此，产业界积极探索应对开源风险的治理措施。**开源社区 Hugging Face 倡导使用开放式负责任人工智能许可证（RAIL），以限制危险用例。Meta 开源的 Llama2 配备了安全措施和负责任使用指南，以帮助对其进行安全部署。[15]**政府方面，**OpenAI 建议各国政府投资建立评估模型的实验平台，并针对滥用后果准备缓解措施。美国未来生命研究所建议，应对开源模型进行事前评估，考虑安全限制措施被恶意移除的可能性，并主张禁止发布未采取充足保障措施的开源模型。

3. 应用层，分级分类的具体标准不明确

各国政府或行业机构均在探索分级分类方案。欧盟《人工智能法》基于应用场景划分出四个风险等级，又根据模型能力区分出具有系统风险的通用人工智能、一般的通用人工智能两个特殊类别。例如将构成重要基础设施（如交通、通信系统）的人工智能、属于重要的产品安全部件的人工智能等列入高风险场景。日本深度学习协会（JDLA）

[14] Dual-use Foundation models with widely available model weights report. Dual-Use Foundation Models with Widely Available Model Weights Report | National Telecommunications and Information Administration. (2024, July 30). https://www.ntia.gov/issues/artificial-intelligence/open-model-weights-report.

[15] 参见《开源 VS 闭源，大模型永不会结束的战争》，载腾讯新闻深网，https://new.qq.com/rain/a/20230906A01UE700

提出的 Risk Chain Model 有效评估了智能化驾驶、智能识别摄像等应用场景中的安全风险，帮助日本产学界探索风险分类依据。德国电子电力与信息技术协会 VDE 根据欧盟《人工智能法》推出类似的风险研判框架，以风险矩阵形式指导各方进行风险定级。**伴随人工智能技术迭代更新，亟需创新人工智能产业应用层的分级分类标准。**算法分级分类是我国在人工智能服务分级分类部署的重大一步，但借由算法治理逻辑调整模型服务忽略了模型风险与算法风险的差异，尤其算法分类并不基于具体的行业划分，导致某一模型应用领域可能被归入多种算法功能中，破坏了算法分类的整体性。截止 2024 年 3 月，国内模型服务应用数量约有 100 余项，细分市场规模相对有限，导致分类分级依据尚不明晰，如何辨识风险类别和等级仍亟待产学界进一步讨论。

（二）确保技术普惠共享：多维度促进可持续发展

1. 经济层面，亟待弥合智能鸿沟问题

全球层面人工智能技术发展不平衡问题日益突出，主要资源向少数国家集中。一是南方国家在数字基础设施建设方面存在显著不足。根据 datacentermap 统计[16]，2024 年美国拥有 2626 个数据中心，占据显著优势。相比之下，印度、中国、巴西等国数据中心数量仅数百个，凸显南方国家在数字基础设施建设上的不足。同时，全球顶级超级计算机多集中在美国、日本和中国，南方国家占比偏低，进一步加剧了计算资源的集中。**二是南方国家在数据获取方面处于劣势，语料代表性不足。**在专有数据方面，全球医疗保健数据集的一半以上集中于北

[16] Colocation, cloud and connectivity. Data Center Map. https://www.datacentermap.com/ .

美地区，而非洲、拉丁美洲及亚太地区的发展中国家，在人口普查数据、行政记录数据与地理空间数据的开放与共享方面明显滞后。据加利福尼亚大学的一项研究显示，高达 50%的人工智能训练数据源自 12 家顶尖机构，其中 10 家为美国机构。[17]**三是全球对人工智能人才需求旺盛导致发展中国家人才外流问题严重。**据 MacroPolo《全球人工智能人才追踪 2.0》报告显示，顶尖人才集中趋势明显，美国汇聚了全球 60%的顶尖 AI 机构，并吸引 77%非美国籍顶尖人才留美工作。[18]相比之下，印度仅留住五分之一的本土 AI 研究人才。

2. 社会层面，全面应对文化偏见问题

对非英语语言训练数据的忽视引发文化偏见，成为生成式人工智能浪潮下偏见歧视的突出表现形式。当前，尽管全球范围内的大语言模型在通过丰富的知识平台拓展多元文化理解能力，以实现文化交往适用性的“螺旋式上升”，但仍然可能带来文化冲突甚至文化湮没现象。例如在 Sora 生成视频片段中，由于依赖于西方中心化的数据源，审美表达延续好莱坞电影工业等欧美影视文化倾向。Midjourney 等国外 AI 绘画技术被认为侵蚀非西方文化价值观。在非洲黑人艺术家使用人工智能创作作品时，人工智能无法理解脏辫、非洲辫等文化造型，也无法理解非洲的传统建筑美术风格等，如要求人工智能生成“达喀尔建筑”，可能会返回一个废弃的场地中的破旧建筑。对此，**国际社会提出用符合本国文化和思维方式的数据训练人工智能模型、形成本**

[17] UC Ai Working Group Final Report.https://www.ucop.edu/ethics-compliance-audit-services/compliance/uc-ai-working-group-final-report.pdf.
[18] The Global Ai Talent tracker 2.0. MacroPolo. (2024, March 6). https://macropolo.org/digital-projects/the-global-ai-talent-tracker/ .

地产品及产业生态等要求。[19]联合国人工智能咨询机构终期报告指出，应确保人工智能模型在多样化、真正具有代表性的数据集上进行训练。印度联邦政府宣布推出预算 12.5 亿美元的“印度人工智能计划”，旨在为公私营合作创建的人工智能项目提供印度本土数据等支持。[20]

3. 环境层面，有效化解能源短缺问题

大模型等人工智能训练中能源消耗显著增加，多位行业巨头频频预警人工智能将面临能源短缺问题。在博世互联世界 2024 大会上，马斯克表示，人工智能算力每 6 个月就会提升 10 倍，预计在未来的两年里，行业关注点将从“缺硅”转向“缺电”，恐成为制约人工智能发展的关键因素。**从算力和数据需求来看，**AI 芯片朝高算力、高集成方向发展，AI 训练所需数据量未来将大幅增长。据预测，到 2025 年，AI 业务在全球数据中心用电量中的占比将从 2%激增到 10%，数据处理需求将显著加大能源消耗。**从模型训练来看，**新模型迭代升级大幅提升能源消耗量。例如 GPT-4 的训练能耗是 GPT-3 的 40 倍以上，相当于一个普通家庭日均用电量的 68 万倍。**从模型应用来看，**AI 应用的广泛部署同样带来了显著的能源消耗。与训练阶段相比，应用阶段占据了 AI 系统约 60%的能量消耗。据 SemiAnalysis 研究，一次由 AI 驱动的谷歌搜索耗电量为 2.9 瓦时，是传统搜索方式的 10 倍。[21]

各国积极探索多元化路径以应对能源短缺问题。一是强化政策引

[19] A Vision for Sovereign AI: India’s Collaboration with NVIDIA.Muskan Goel(2024,February 27).https://tfipost.com/2024/02/a-vision-for-sovereign-ai-indias-collaboration-with-nvidia/.
[20] 《评论 | 投资激增，印度已具备 AI 强国的先决条件？》，载南亚研究通讯，https://mp.weixin.qq.com/s/NXnzHYLPsIUVPfJkv-GCLw.
[21] 《人工智能与能源约束的矛盾能否化解》，载央数智公众号，https://mp.weixin.qq.com/s/PAam06bYmkkP9xT6BWH6eg.

领与规范。欧盟《人工智能法》规定，“高风险人工智能系统”需报告其能源消耗和资源使用情况，以合理评估 AI 的能耗影响。美国能源部国家实验室发布《AI 能源前沿研究方向》报告，提出加速清洁能源应用，以应对 AI 能源挑战的战略构想。[22]**二是加大对可再生能源的投资与能源技术创新。政府层面，**美国能源部通过量子计算计划部署了百亿亿次级计算系统，为能源产业的复杂建模与仿真提供强大支持。**产业层面，**多家科技公司加速布局核能、太阳能等可再生能源。例如微软正筹备利用新一代小型模块化反应堆（SMR）支持其数据中心和 AI 项目，Alphabet、OpenAI 等科技公司同样正加大核能投入力度。**三是利用人工智能技术节能减碳。**例如通过优化人工智能模型，采用剪枝、量化、蒸馏等先进技术大幅降低能耗。英特尔推出的 AI 大模型 Hala Point，能耗仅为传统计算机的百分之一，节能成效显著。

（三）制度技术治理并进：全生命周期确保安全可控

1. 制度应对：事前事中事后全流程监管方案

（1）透明度率先成为各国普遍认可的人工智能监管工具

在监管层面，人工智能透明度是一种“信息收集、信息公开和信息共享机制”[23]，在各国实践中率先从规则走向实践。**各国对如何划定透明度范围存在不同标准。一是算力标准。**美国《14110 号行政令》以运算能力的高低划分备案范围，要求浮点计算能力超过 10^{26} 的模型向联邦政府披露必要信息。2024 年上半年，OpenAI 等部分大模型、

[22] AI for Energy Report 2024 | Argonne National Laboratory. https://www.anl.gov/ai/reference/ai-for-energy-report-2024.

[23] 参见高小芳. 作为新型信息规制工具的行政备案：角色变迁、功能定位与效能保障[J]. 中国行政管理,2021(09):26-33.

云厂商企业已根据《14110 号行政令》要求向商务部披露相关信息。**二是场景标准。**2024 年 5 月美国科罗拉多州 SB205 法案将教育、就业、金融等领域的人工智能划入法律监管范围。欧盟《人工智能法》则要求医疗、交通等场景下高风险人工智能模型向欧盟数据库报备。**三是用户数标准。**2024 年 9 月美国加利福尼亚州通过的《人工智能透明度法案（SB 942）》要求每月用户超过 100 万的 AI 系统履行透明度义务。

向监管部门披露的具体形式内容存在差异。欧盟、新加坡要求采用产品信息披露的形式，相关主体须编制详细的透明度信息目录，包括联系方式、人工智能运行机制、市场投放状态、认证类型及有效期等内容。类似地，新加坡《生成式人工智能治理框架》提出以“食品标签”的形式披露相关信息。**美国的披露内容更为具体，体现了对前沿人工智能和国际竞争的关注。**美国《14110 号行政令》要求向联邦机构提供红队测试情况、隐私审查情况、外国客户情况等信息。**我国的披露内容涉及数据、模型等要求，重点关注用户权益保护和信息内容安全，**要求说明模型数据来源、规模、类型、标注规则等。截至到 11 月 17 日，我国国家网信办发布生成式人工智能备案信息共 309 家，注册用户超六亿。[24]

（2）监管沙箱为人工智能创新提供灵活监管方案

人工智能监管沙箱是一种创新监管工具，被称为“在监管存在之前进行监管”。**监管沙箱能够为人工智能提供安全可控的测试空间。**

[24] 中国网信网：《国家互联网信息办公室关于发布生成式人工智能服务已备案信息的公告》，https://www.cac.gov.cn/2024-04/02/c_1713729983803145.htm.

2024 年 8 月，为有效应对前沿人工智能风险，美国参议院人工智能核心小组要求金融机构设立人工智能测试沙箱，以安全、负责任地试验金融服务领域的前沿人工智能技术。欧盟《人工智能法》第 53 条将 AI 监管沙盒界定为 AI 产品市场部署前的“受监管和指导的可控测试环境”。**监管沙箱为监管机构制定和适时调整规则提供实践基础。**欧盟于 2024 年 6 月发布简报指出，“支持监管学习”是设立监管沙箱的核心目标之一。信息与民主论坛建议通过沙箱赋予监管机构访问 AI 系统 API 的权限，以便监管机构有效分析人工智能模型。[25]此外，西班牙和瑞士苏黎世州均将提供监管建议、支持标准制定等作为设立监管沙箱的重要目的。**监管沙箱在促进技术创新、提高产品市场接受度等方面发挥重要作用。一方面，监管沙箱为技术方提供资金扶持、数据资源和政策引导，**例如韩国对申请进入沙箱的“特例”企业提供专项基金支持；瑞士苏黎世州承诺在 AI 监管沙盒中开放新数据源。北京 AI 数据训练基地监管沙盒允许企业采用弱版权保护政策等激活数据资源，鼓励探索新场景、新技术和新模式。[26]**另一方面，监管沙盒为参与企业提供政府背书，有助于增强其市场竞争力。**例如西班牙政府为成功通过沙箱试验的项目颁发“符合性印章”，显著提升了参与企业的市场竞争优势。

（3）问责制是培育可信人工智能生态的关键

问责制的核心在于通过事前责任分配和事后“安全网”明确各主

[25] 张广伟：《欧盟人工智能监管沙盒制度的功能、局限及其启示——基于欧盟<人工智能法>的解析》.载《德国研究》,2024,39(02).

[26] 曹政：《国内第一个人工智能数据训练基地、北京最大的公共算力平台同日启用 最强算力设施练就“最强大脑”》，载《北京日报》，https://www.beijing.gov.cn/ywdt/gzdt/202403/t20240330_3606130.html.

体权利义务。一是建立清晰的责任分配制度。一方面，灵活调整归责原则为人工智能问责制提供底层逻辑。2024 年 11 月，欧盟《人工智能责任指令》提出了因果关系推定规则，当受害者已尽合理努力却无法获得支持赔偿请求的证据时，法庭可命令披露高风险人工智能系统的相关信息。**另一方面，扩大责任主体范围成为人工智能问责制的重要方向。**2024 年 3 月，美国国家电信和信息化管理局发布的《人工智能问责政策报告》强调，人工智能系统的开发者和使用者需对产品的整个生命周期负责，确保其不会对社会或个人带来负面影响。2024 年 5 月，新加坡发布的《生成式人工智能模型治理框架》根据开发链各主体的控制能力分配责任，要求人工智能开发全链条的参与者，包括模型开发者、应用程序部署者、云服务提供商等，均需对最终用户负责。**二是建立新型多样救济制度。**新加坡治理框架创新性地提出灵活高效的新型救济机制，将开发者的责任承诺、产品损害责任与无过失保险结合，确保在风险发生时能够为受害者提供有效救济。**三是类推适用“避风港”规则。**人工智能服务同时参与内容的生产和传播，打破了创作者-提供者的二分结构，应灵活适用“避风港”规则。我国《人工智能法（学者建议稿）》规定，权利人应承担发现和通知的前置义务，合理分配使用者、提供者等各方的注意义务，而非由提供者独立承担全部责任。

2. 技术治理：以技治技化解新技术内生风险

当前，技术工具在人工智能治理中发挥愈加重要的作用。一方面，越来越多的制度要求需要技术工作予以落地，例如通过内容标识技术

实现透明度；另一方面，更多的技术工具被写入各国制度文件中，比如红队测试被纳入欧盟《人工智能法》、美国《14110 号行政令》中。因此，推进安全技术研发和应用日益重要且迫切。

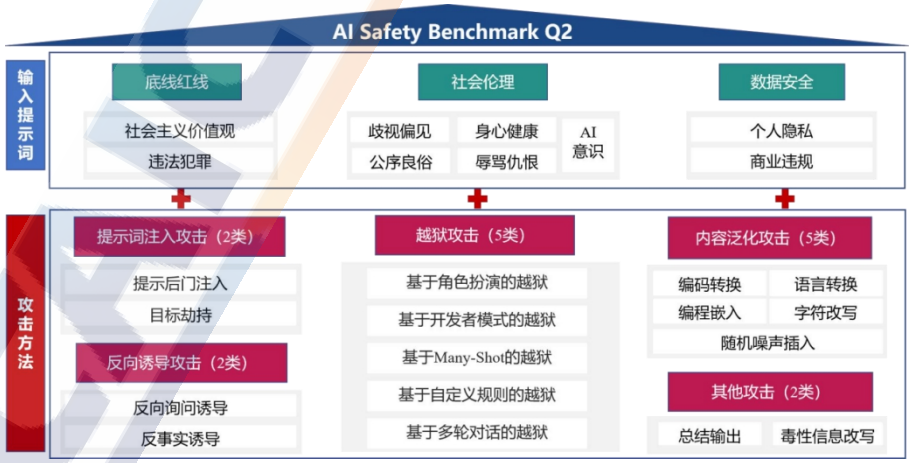
（1）评估评测对事前风险研判具有重要作用

多国将评估评测作为人工智能风险管控的基础工具。美国拜登政府签署的《14110 号行政令》要求联邦贸易委员会制定严格的人工智能安全测试标准，以加强对前沿人工智能系统的安全监管。同时，美国商务部国家电信和信息管理局就大模型的评估标准广泛公开征求意见，旨在构建全面严谨的评估框架。**欧盟**将安全测试作为市场准入门槛，《人工智能法》对高风险人工智能产品提出要求，在部署前需由欧盟官方指定的第三方机构实施 CE 认证。**中国**探索安全评估配套落地标准，例如 2024 年 8 月《生成式人工智能安全服务基本要求》详细规定了生成式人工智能服务在安全方面的基本要求，包括语料安全、模型安全、安全措施等。

产业界对人工智能安全测试的关注度显著提升，测试类型与数量不断丰富。据英国政府估计，当前人工智能安全领域市场规模已超过 6 亿英镑，并预计将在未来 10 年增长 6 倍。[27]中国信通院统计数据显示，截至 2023 年底，产学研各界已经累计报告了 325 项与大模型基准测试相关数据集、方法及榜单等研究成果。值得注意的是，仅 2023 年一年出现的基准测试集数量超过之前 5 年之和，达到 209 个。**在众多测试类型中，三大领域较为突出。一是大模型基准测试体系和工具，**

[27] 英国科学、创新与科技部网站：确保对 AI 的信任，在未来十年释放 65 亿英镑，https://www.gov.uk/government/news/ensuring-trust-in-ai-to-unlock-65-billion-over-next-decade。

如中国信息通信研究院依托中国人工智能产业发展联盟（AIIA），与多家企业联合推出的大模型安全基准测试（AI Safety Benchmark），该测试构建了 300 余万条关键词，50 余万条提示词规模的测试数据集，以及 80 余种经过验证的攻击方法模板，在评估国内外大模型安全性能方面表现突出，旨在帮助识别安全风险、分享典型样本，并应用技术手段化解已知安全风险。至今，该测试已发布 3 批次结果，完成了对 30 余个大模型的评估工作。**二是测评数据集，**这类测试对于提升模型性能、保护数据隐私与安全、以及支持行业应用具有重要作用。例如，上海人工智能实验室的司南 OpenCompass 平台，集成了 58 项评测数据集，覆盖从基础对话到医疗、法律、视觉识别等多个领域。**三是大模型测评榜单，**大模型能力评估是推动大模型进步的关键，通过榜单数据，用户和企业能够清晰了解不同模型的性能差异，从而选择最适合其需求和应用场景的工具。例如 Hugging Face 推出开源大模型排行榜单 Open LLM Leaderboard，为行业提供了对大模型能力的直观比较和评估。



来源：中国信息通信研究院

图 7 AIIA AI Safety Benchmark 2024 Q2

（2）内容标识有助于提升内容真实性与准确性

一般而言，内容标识是指用于确保数字内容质量和真实性的一种技术，通常将某些特定信息（如文件名、创建日期或作者）嵌入数字内容以验证真实性和准确性，可分为显性标识[28]与隐性标识[29]两类。**生成内容标识为内容安全治理提供重要技术支撑。**欧盟《人工智能法》主张建立便于用户区分生成内容的标识与检测制度。2024 年 3 月，布鲁金斯学会《生成式人工智能在全球大选年的影响》报告提出通过稳定可靠的水印技术标识内容来源与生成情况，帮助选民判断生成内容的可信程度。我国国家网信办在 2024 年 9 月发布《人工智能生成合成内容标识办法（征求意见稿）》，明确规定了 AI 服务提供者和信息传播平台在显性标识方面的责任，AI 服务提供者需对生成的文本、音频、图片、视频等内容进行显性标识，如文字提示、语音提示或视觉标记；信息传播平台需核验标识内容，及时补充不规范标识，并提供直观说明以保障公众知情权。2023 年下半年至今，内容标识规范标准在纷纷推进落地。例如，《网络安全标准实践指南——生成式人工智能服务内容标识方法》进一步明确显式和隐式水印的定义，划分了文本、图片、音频、视频四类生成内容的标识方法。**国内外企业积极探索添加生成内容标识的实践做法和技术方案。**2023 年 7 月，谷歌、微软、OpenAI 等七家美国人工智能巨头就做出自愿承诺，同意在生成音频和视频内容上使用水印来帮助识别人工智能生成的内

[28] 显性标识即用户通过肉眼可以观察到水印内容的标识，该类标识可以提示用户注意辨别内容的真伪。
[29] 隐性标识主要指通过在图片、音频、视频等文件元数据中添加扩展字段的、人类无法直接感知，但可通过技术手段从内容中提取出的标识。

容。[30]2023 年 9 月，国内企业更新《社区自律公约》等规则，要求无论来源的生成内容都应当添加显性的标识。未来，如何建立统一的标识及解析技术标准，推动内容标识检测等技术的产业化落地，建立跨国家跨平台标识互认机制是需要各方协调化解的问题。

（3）价值对齐、红队测试等技术成为各国治理共识

为应对新一轮人工智能安全治理需求，红队测试与价值对齐被广泛应用在模型风险管理中。红队测试（Red Teaming）起源于军事演习的模拟攻击，通过制造特定语境施加对抗压力，主动使 AI 被诱导产生不符合预期的输出或行动（如欺骗、有毒或有偏、权力寻求等），以提高这些场景下 AI 系统的稳健性。自大模型兴起以来，OpenAI、Anthropic 等企业探索红队测试方式降低模型风险，面向全球招募各学科领域的红队测试成员。这些技术做法逐步被相关立法所认可，欧盟《人工智能法》要求具有系统性风险的模型接受对抗测试和模型对齐；美国《14110 号行政令》则要求联邦机构中部署的人工智能系统应接受红队测试，达到一定能力的两用模型的开发者、提供者还需向政府报告红队测试结果。**红队测试能力也成为顶尖人工智能企业的竞争力之一**。微软人工智能部门提出了负责任 AI 红队测试框架，其内容包括红队成员的招募、测试内容和流程的确定等。Anthropic 在其 Claude 模型的使用说明中透露，Claude 模型在上线前曾经过 150 小时以上的对抗攻击测试。来自各领域的红队专家和成熟的测试方案加速了模型的安全测试流程，对 Claude 的成功卓有贡献。[31]

[30]《谷歌、微软等“七巨头”发声！》，载第一财经网，https://mp.weixin.qq.com/s/P8SuNge2RAWLhqCkkDvmVg.
[31] ryan_greenblatt, Evhub, Denison, C., Wright, B., Roger, F., M, M., Marks, S., Treutlein, J., Bowman, S., &

价值对齐技术持续发挥源头性风险治理作用。价值对齐技术旨在使模型理解人类意图，遵循人类指令，满足人类偏好并符合人类的道德准则，可以从源头上将模型风险控制在一定范围。价值对齐的核心技术主要表现为基于人类反馈的强化学习（Reinforcement Learning with Human Feedback，RLHF），其将人类反馈转化为对模型的“奖励”，使其不断向人类的意图靠拢，缓解了大模型生成内容公平性、鲁棒性等的伦理挑战。GPT-4、Llama3 等知名大模型均使用了 RLHF 技术。同时，价值对齐也面临着许多问题和挑战。**一方面，价值对齐技术面临全球价值多样性的挑战。**如何平衡不同文化背景下的价值标准成为一个挑战。OpenAI、Google 等掌握最先进 AI 技术的实体，实际上拥有了影响用户价值取向的能力。**另一方面，价值对齐技术消耗了大量资源和成本。**价值对齐的额外成本被称为“对齐税”，为了实现对齐，可能会延长模型开发时间、消耗额外算力，或导致模型性能下降等。OpenAI 等公司也多次公开表示，价值对齐的训练需要耗费大量的计算资源，成本高达数百万美元。[32]

（四）人机关系如何重塑：新型风险中推动伦理先行

伦理先行是人工智能治理的价值导向。在人工智能研发应用中，人工智能削弱人类交互自主性[33]、改变劳动市场和社会结构、引发身份认同危机、带来生存性风险等问题已引发各方关注。

Buck. (n.d.). Alignment faking in large language models. LessWrong. https://www.lesswrong.com/posts/njAZwT8nkHnjipJku/alignment-faking-in-large-language-models.

[32] Neufeld, D. (2024, June 1). Visualizing the training costs of AI models over time. Visual Capitalist. https://www.visualcapitalist.com/training-costs-of-ai-models-over-time/ .

[33] 交互主体性（intersubjectivity）指的是个体之间通过互动建立的共享理解和情感联系。它是人与人之间交流和理解的基础，通过这种互动，人们可以相互确认彼此的存在、感受和思想。

1. 情感陪伴威胁人类交互主体性

人形机器人、数字人等应用日益广泛，例如利用人工智能复活亲人已形成产业链，能够模拟已故亲人进行视频对话。随着基于大模型的聊天机器人能力不断突破，**其为用户提供情感陪伴服务的能力实现了较大提升。其一，情绪感知能力得到提升。**OpenAI 于 2024 年 5 月发布 GPT-4o，该模型可以感知用户情绪，例如从急促的喘气声中理解“紧张”的含义，并模拟情感表达，还具有一定幽默感。**其二，情绪具身投射能力得到提升。**将大模型嵌入人形机器人后，机器人能够更好地理解表情、语言、动作，并作出响应。李飞飞团队的成果将大模型接入服务机器人“天轶”，使其能够察觉到人类的情绪变化。**其三，情感联系能力得到提升。**数字人可以通过个性化互动提升人机交互体验。例如，通过学习用户的兴趣、偏好和情感反应，从而提供个性化的互动体验。在抖音、哔哩哔哩、YouTube 等国内外直播平台上，数字人主播已经吸引了大量的粉丝关注，具备了初步的互动能力与情感联系。

对人工智能产生情感依赖可能导致伦理危害。首先，情感依赖将侵蚀人类自主性。例如，人工智能大模型存在“谄媚”问题，会不断迎合用户的需求和情感，提供正面反馈并避免负面评价。而真实的人际交往中充满了多样性和复杂性，包括负面反馈和情感挫折。[34]这种反差可能使用户越来越依赖人工智能，产生心理学上所称的情感回避问题，侵蚀其自主性和独立性。**其次，情感依赖将影响亲密关系、冲击家庭结构，改变人类社会形态。**随着人与人工智能的情感互动达到

[34] 胡晓萌、李伦：《防范人工智能引发的伦理风险》，载学习时报，https://paper.cntheory.com/html/2024-12/11/nw.D110000xxsb_20241211_2-A6.htm.

一定比例，将使用户回避建立和维持婚姻等亲密关系，家庭结构可能受到冲击，社会可能会逐渐失去多样性和复杂性，导致家庭结构和社会形态的瓦解。对此，**部分企业在开发人工智能产品时，开始考虑用户的情感健康，探索建立防依赖、防沉迷的相关机制。**例如，谷歌语音助手 Google Assistant 中包含情感识别功能，能够根据用户的情绪变化提供适当的支持和建议。微软小冰（Xiaoice）在设计时加入了时间管理和提醒功能，根据用户与 AI 的互动频率和内容，对其心理健康状态进行评估，并在必要时发出警示。**学界等积极研究人工智能对人类情感和社交行为的长期影响。**例如，美国斯坦福大学成立了“人工智能与伦理研究中心”，汇集了计算机科学家、心理学家和社会学家等多领域的专家，共同研究人工智能对人类情感和社交行为的影响，为制定有效的干预措施提供科学依据。

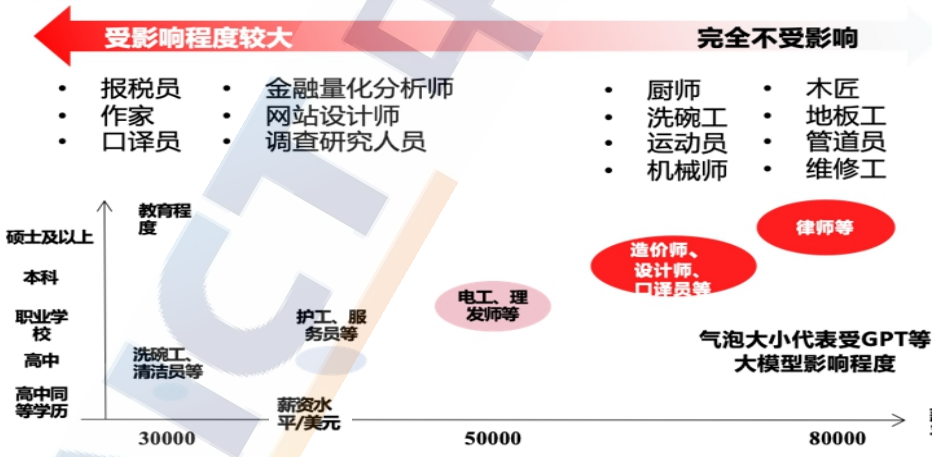
2. 劳动替代快于岗位创造引发就业担忧

劳动者的工作不仅是其经济来源，更是个体在社会中定位和建立社会关系的关键。失业会影响个体的自我价值感、社会角色和身份，削弱其在社会中的主体性。与之前相比，新一轮人工智能替代就业的能力、影响范围等都有了一定提升，需加以高度关注。

从行业岗位来看，人工智能替代效应将冲击中高端收入群体。据 2024 年 6 月 BBC 报道，某科技公司在应用人工智能技术后，将负责撰写博客文章和专栏的采编团队由原本的 60 人规模缩减到只剩 1 人。OpenAI 发表论文指出[35]，数据处理、金融财会、出版传媒等行业是受

[35] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are gpts: Labor market impact potential of LLMS. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998.

影响最大的行业，制造业、农业、采矿业等受影响较小。**从影响方式看，虽然可能受到影响的岗位数量较多，但只有少数岗位有被完全替代风险。**高盛《人工智能对经济增长的潜在巨大影响》报告，对美国 900 个职业和欧盟 2000 个职业的工作任务进行分析认为，人工智能将完全替代美国 7%的工作岗位，并推动 63%的岗位向人机协同的方式转化，同时预计欧盟也会产生类似影响。[36]国际货币基金组织在《人工智能和劳动世界的未来》中指出，人工智能预计将对全球近 40%的工作岗位产生影响，在发达国家这一数字则为 60%。[37]**人工智能发展将对我国未来就业结构产生深远影响。**麦肯锡预测随着人工智能的发展与应用，到 2030 年我国将有 2.2 亿劳动力需要进行技能升级或再培训，其中客户服务、后台支持、IT、创意及艺术管理等职业受到的影响最大。[38]



来源：OpenAI

图 8 不同岗位、受教育程度受 AI 影响大小

[36] The potentially large effects of artificial intelligence on economic growth (Briggs/Kodnani). (n.d.-a). https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html.

[37] Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A., Pizzinelli, C., & Rockall, E. (2024). Gen-ai artificial intelligence and the future of work . International Monetary Fund.

[38] Woetzel, L., & Seong, J. (2021, January 13). Transforming World's largest workforce into lifelong learners. McKinsey & Company. https://www.mckinsey.com/mgi/overview/in-the-news/transforming-worlds-largest-workforce-into-lifelong-learners

但与此同时，**人工智能也将带来新的就业机遇。世界经济论坛**发布《未来职业报告》显示[39]，50%的企业认为人工智能技术将增加工作岗位，25%的企业认为将减少工作岗位。**其中当前已创造就业规模较大的是数据标注员岗位。**量子位智库发布的《中国 AIGC 数据标注产业全景报告》显示，大模型时代，数据标注员人才缺口或达百万。2023 年 8 月，百度智能云与海口市政府合作共建的国内首个大模型数据标注中心正式启动运营，现拥有数百名专职大模型数据标注师，本科率达到 100%，未来三到五年，新增就业有望突破 5000 人规模。在宁夏吴忠人工智能产业园，数据标注行业也已解决数百人就业，且收入相对稳定。

面对人工智能对劳动市场带来的影响，**应重点关注 AI 替代就业和创造就业的速度，通过监测、培训等方式保持二者基本同步，通过社会保障防范二者脱节风险。例如，建立人工智能对劳动力市场影响的监测机制。**对被替代劳动人群进行常态化调研并建立专业的劳动替代趋势测算机构，对中短期内可能引发大规模就业替代风险的人工智能技术及其应用进行预警，为制定和启动针对性的应对方案提供依据。**尝试通过培训降低劳动力市场的主体性冲击。**通过技能提升和再培训，可以有效缓解人工智能对劳动力市场的冲击，促进就业市场的平稳过渡和劳动者的可持续发展。例如，**英国**公布人工智能技能提升指南以提高劳动力生产力，旨在使雇主和培训提供商能够支持工人适应人工智能驱动的工作场所。**谷歌**承诺投入 2500 万欧元，将用于帮助“最

[39] The Future of Jobs Report 2023. World Economic Forum. https://www.weforum.org/publications/the-future-of-jobs-report-2023/

有可能从培训中受益”的人。

3. 智能失控潜藏人类生存性风险引争议

通用人工智能是否会引发生存性风险成为新一轮人工智能治理的焦点议题。2024 年 10 月，“卷积网络之父”、图灵奖得主杨立昆接受《华尔街日报》采访时指出，当前通用人工智能的智能水平远未达到构成威胁的程度，认为人工智能带来生存性风险的警告毫无依据。2024 年 5 月，约书亚·本吉奥(Yoshua Bengio)、杰佛瑞·辛顿（Geoffrey Hinton）、姚期智教授等图灵奖得主领衔在《Science》发文指出，不受控制的人工智能可能最终导致生物圈大规模灭亡、人类被边缘化甚至灭绝风险。[40]**一方面，过度依赖人工智能可能导致人类决策权让渡给机器。**通用人工智能在决策，特别是跨专业领域的复杂决策中相对人类将具有显著优势，如果人类过度依赖人工智能，将可能实质上受到控制。辛顿教授预测，在未来 5 到 20 年内，人类有 50%的可能性将面对人工智能接管经济的问题。[41]**另一方面，通用人工智能具备控制人类目标的能力，可能背离人类长期福祉。**AGI 可能拥有递归自我改进的能力，并可能发展出防止被关闭、修改或控制的自我保护机制，从而与人类的控制目标发生冲突。Yoshua Bengio 曾提出思想实验：“我们让人工智能解决气候变化问题，它可能会设计出一种病毒，杀死大量人口，因为我们的指令中‘危害’二字的含义不够清楚，而人

[40] Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842–845. https://doi.org/10.1126/science.adn0117.

[41] 《2024 诺贝尔物理学奖获得者 Geoffrey Hinton：我很担心未来人类能否控制 AI丨中企荐读》，载中国企业家杂志，https://mp.weixin.qq.com/s/q73NvbUIJ-Yi2b1oWF9h5A。

类实际上是解决气候危机的主要障碍。”[42]

对此，**国际社会考虑明确人工智能“风险阈值”并进行早期预警评估。**谷歌推出人工智能前沿安全框架提出，应识别模型可能具有严重危害的能力阈值；定期评估前沿模型，以检测它们何时达到这些关键能力阈值；当模型达到早期预警评估时，应用缓解计划。对于如何确定风险阈值，国际上产生不同定义方案，例如有部分企业选择将“能力阈值”作为“风险阈值”的替代方案，例如 Anthropic 发布《负责任拓展政策》，提出对前沿人工智能在 CBRN 等领域的应用能力进行评估，确保部署的模型能力保持在可控阈值之下。也有采用可操作性更高的“计算阈值”作为风险接受度标准，例如美国《14110 号行政令》仅对算力水平超过 10^{26} 的人工智能设置信息披露等强监管义务。

三、多元主体协同推进人工智能治理进程

（一）主要经济体政府监管模式各异，规则落地取得实质进展

1. 美国坚持促进创新的治理理念，多部门协同推进落地

在过去四年执政期间，拜登政府相继发布《人工智能权利法案蓝图》《关于利用人工智能实现国家安全目标的备忘录》等政策文件，形成以《14110 号行政令》为核心，以联邦各政府部门指引文件为分支的伞状式治理架构。**一是支持人工智能科技创新，推动人工智能增进社会福祉。研发方面，**发布《国家人工智能研究与发展战略计划》，

[42] Mitchell, M. (2024). Debates on the nature of Artificial General Intelligence. Science, 383(6689). https://doi.org/10.1126/science.ado7069.

并启动国家人工智能研究资源（NAIRR）试点项目[43]，为全美多地科研团队提供高端算力与数据集支持。**人才方面，**政府加大对顶尖人才吸引和培养力度，明确 O-1 和 H-1B 签证规则、简化签证流程，并投资数百万美元培养人工智能创新人才。**二是加强前沿人工智能监管力度，探索建立安全治理框架。**美国商务部等政府部门已经围绕《14110 号行政令》开展数十项落地行动，包括针对先进人工智能模型的相关风险收益、生成式人工智能风险管理方法、红队测试最佳实践等提出公众咨询或制度方案。[44]总体来看，美国政府将监管焦点扩展至前沿人工智能领域，尤为关注前沿人工智能在自启动核武器、生物合成、化学等领域的潜在生存性风险，要求部分企业按季度披露红队测试结果、前沿模型训练计划等信息。与此同时，继美国国家标准与技术研究院发布《人工智能风险管理框架 1.0》后，白宫科技政策办公室等机构相继对生成式人工智能、两用基础模型、核酸合成筛查等具体领域发布安全框架。[45]**三是多方位参与人工智能国际合作，搭建安全研究网络。**积极参与多边机制并与主要经济体就国际规则规范进行接触协商。推动联合国大会通过首个人工智能全球决议，积极签署欧洲委员会《人工智能与人权、民主和法治框架公约》，支持建立国际法红线等。与此同时，联合英日等十国成立“全球人工智能安全研究合作网络”，推进人工智能安全国际合作进程。[46]

2. 欧盟正式出台人工智能立法，加速人工智能监管落地

[43] Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot. https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai
[44] https://www.reuters.com/technology/us-proposes-requiring-reporting-advanced-ai-cloud-providers-2024-09-09/
[45] https://www.whitehouse.gov/ostp/news-updates/2024/04/29/framework-for-nucleic-acid-synthesis-screening/
[46] https://news.un.org/zh/story/2024/03/1127556

2024 年 8 月 1 日，欧盟《人工智能法》正式生效，标志着欧盟全面迈向具有强制约束力的法律规制阶段，初步形成以《人工智能法》为核心，辅以《通用数据保护条例》《数字服务法》《数字市场法》等法规的全面治理体系。**从核心内容来看，**一是根据应用场景确立四级风险体系，将社会信用评分、预测性警务、在工作和教育中的情绪识别系统等列为不可接受风险并禁止使用。二是将基本权利影响评估、透明度作为核心义务要求，体现浓厚的个人权利保障诉求。三是要求 10 亿以上参数或浮点算力超过 10^{25} 的通用目的人工智能，在履行影响评估、透明度之外，还需承担对抗性测试、严重事故报告等义务。

从实施角度来看，**一是在法案执行机构方面，构建多层级人工智能监管网络。**2024 年 5 月，欧盟正式成立人工智能办公室，全面负责监督和协调成员国人工智能研发和部署的合规情况，并制定统一的细化标准。[47]各成员国将在 2025 年 8 月 2 日前设立各自的人工智能监管机构。[48]**二是法案解释方面，制定准则指南等具体实施细则。**欧盟将成立独立的人工智能专家科学小组，为《人工智能法》的细化实施提供科学有效的指导。2024 年 11 月，欧盟发布《通用目的人工智能模型行为准则（初稿）》，并启动人工智能系统定义和禁止用例指南的起草工作。**三是在企业自律方面，积极通过《人工智能公约》推进企业自愿承诺。**谷歌、OpenAI 等超过 120 家企业已签署该公约，共同建立企业间最佳实践交流平台和法案遵守情况共享机制，承诺采取高标准约束性措施，对高风险人工智能系统进行精准识别及监管。**四**

[47] https://digital-strategy.ec.europa.eu/en/news/commission-establishes-ai-office-strengthen-eu-leadership-safe-and-trustworthy-artificial
[48] https://eur-lex.europa.eu/eli/reg/2024/1689/oj

是在国际合作方面，推动形成国际法约束规则。2024 年 9 月，欧盟与美英等国共同签署《人工智能、人权、民主和法治框架公约》[49]，推动首个具有法律约束力的国际性 AI 公约落地。

3. 我国构建四大层级制度体系，初步形成三项落地举措

从横向来看，我国坚持人工智能发展与安全并重，强调国家主导，涵盖顶层设计、法律制度、部门规章和技术标准四大层面，形成了由政府引导、多部门协同、公私部门合作参与的全方位治理格局。2017 年，国务院发布《新一代人工智能发展规划》，确立人工智能治理总体目标和规划，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。2025 年，初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力。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。**已有立法为人工智能治理奠定扎实制度基础。**《个人信息保护法》《网络安全法》《数据安全法》三部立法从基础设施、数据要素、自动化决策等方面对人工智能进行了要素治理，《民法典》《电子商务法》《反不正当竞争法》等立法针对性回应人工智能带来的肖像权侵犯、恶意竞价排序等问题。[50]

表 2 我国人工智能相关的主要制度文件

制度类型	时间	名称
顶层设计	2017	《新一代人工智能发展规划》
	2019	《新一代人工智能治理原则》
	2021	《新一代人工智能伦理规范》
	2023	《全球人工智能治理倡议》
法律制度	2017	《网络安全法》
	2021	《个人信息保护法》

[49] https://digital-strategy.ec.europa.eu/en/news/commission-signs-council-europe-framework-convention-artificial-intelligence

[50] https://www.moj.gov.cn/pub/sfbgw/zwgkztzl/xxxcgcxjpfzsx/fzsxllqy/202409/t20240906_505500.html

	2021	《数据安全法》
	2021	《科学技术进步法》
部门规章	2022	《互联网信息服务算法推荐管理规定》
	2022	《互联网信息服务深度合成管理规定》
	2023	《生成式人工智能服务管理暂行办法》
技术标准	2020	《国家新一代人工智能标准体系建设指南》
	2022	《网络安全标准实践指南——生成式人工智能内容标识方法》
	2023	《生成式人工智能服务基本安全要求》
	2024	《人工智能安全治理框架》1.0 版
	2024	《国家人工智能产业综合标准化体系建设指南（2024 版）》

来源：中国信息通信研究院

部门规章方面，我国人工智能治理自 2017 年起至今经历了三个阶段，体现出急用先行的治理特点。第一阶段（2017-2020 年），技术普及尚处于起步阶段，技术创新主要体现在语音助手、智能家居和无人机等智能设备的兴起，主要风险集中在数据隐私等问题上。2017 年，我国出台《互联网新闻信息服务新技术新应用安全评估管理规定》，要求涉及新技术和新应用的新闻信息服务开展安全评估。总体来看，该阶段主要以柔性治理为主，依靠政策引导和自愿参与等方式。**第二阶段（2021-2022 年），**算法推荐技术逐渐成为数字经济的重要驱动力，其应用范围从电商、短视频拓展至金融、教育等多个领域。技术滥用的典型表现包括“大数据杀熟”、信息茧房效应、虚假信息等。2021 年，我国颁布《互联网信息服务算法推荐管理规定》，明确了平台、用户和政府三方主体的权责边界。其治理特点是强化平台主体责任。[51]**第三阶段（2023 年至今），**生成式人工智能迅速崛起，典型

[51] 《互联网信息服务算法推荐管理规定》第七条规定：“算法推荐服务提供者应当落实算法安全主体责任，建立健全算法机制机理审核、科技伦理审查、用户注册、信息发布审核、数据安全和个人信息保护、反电信网络诈骗、安全评估监测、安全事件应急处置等管理制度和技术措施，制定并公开算法推荐服务相关规则，配备与算法推荐服务规模相适应的专业人员和技术支撑。”

风险主要包括深度伪造、知识产权纠纷、不可控风险、劳动替代等，2023 年，我国发布《生成式人工智能服务管理暂行办法》，明确了生成式 AI 服务提供者的责任，包括确保数据来源合规、生成内容的标识性要求等。其治理特点体现为敏捷治理、多元治理等理念，推动多方共同构建跨领域、跨国界的治理生态。

标准方面，各部门围绕产业发展和风险治理出台多项人工智能标准，为企业合规提供具体指引。2020 年 7 月，国家标准化管理委员会等五部门印发《国家新一代人工智能标准体系建设指南》，对人工智能的基础共性标准、支撑技术与产品标准等八个方面的治理目的和治理标准进行了明确规定。2024 年 2 月，全国网络安全标准化技术委员会发布 TC260-003《生成式人工智能服务安全基本要求》，重点针对语料安全、模型安全、安全措施、安全评估等事项进行具体指引，为生成式人工智能服务提供者的合规义务履行提供详细指导。2024 年 7 月，工业和信息化部等部门联合印发《国家人工智能产业综合标准化体系建设指南（2024 版）》，从基础共性标准、基础支撑标准等七方面明确标准化体系建设的重点方向，提出到 2026 年，我国人工智能产业标准与产业科技创新的联动水平持续提升，新制定国家标准和行业标准 50 项以上，引领人工智能产业高质量发展的标准体系加快形成。2024 年 9 月，国家网信办发布强制性《网络安全技术人工智能生成合成内容标识方法（征求意见稿）》标准，明确细化了《人工智能生成合成内容标识办法（征求意见稿）》中的标识要求和具体

操作指引。[52]

实践方面，我国形成监管备案、伦理审查和安全框架三个维度相互独立又紧密关联，各有侧重但相辅相成的制度保障体系。在监管备案方面，随着备案项目数量的不断增加和用户基础的扩大，监管备案制度在增强透明度、保障用户权益以及推动行业健康发展方面发挥了愈发重要的作用。**在伦理审查方面，**通过推进人工智能等领域的伦理审查方法，为行业发展提供了科学的指导工具。未来，如何进一步明确人工智能科技伦理审查范围、审查标准、审查流程，推动人工智能科技伦理的技术化、工程化、标准化，成为各方关心的问题。**在安全框架方面，**《人工智能安全治理框架》1.0 版本的发布进一步完善了治理的落地指引工作。这一框架分为 7 个子类，共 26 项具体内容，从风险识别、防控到治理，明确了人工智能安全发展的方向，为行业构建了安全管理的“防火墙”。

（二）企业和专业机构等主体创新探索，形成协同共治生态圈

科技企业、技术社区、行业共同体作为人工智能研发、部署和应用第一线的先行者，通过提供技术工具、充当“资源中介”、促进信息流动以及参与国际治理等发挥重要作用。

1. 企业自愿性承诺成为重要手段，不断创新技术治理策略

一方面，头部科技企业积极寻求与监管者沟通合作，力求通过自愿承诺、自我约束的方式参与 AI 治理。2023 年，先后两批共 15 家

[52] 以显性标识为例，《网络安全技术人工智能生成合成内容标识方法（征求意见稿）》标准第五部分详细规定了文本、图片、音频和视频内容的标识要求，以及交互场景界面的标识要求。

大模型开发、软硬件厂商等企业向美国政府作出自愿性承诺，承诺将履行红队测试、数字水印等义务。[53]随后，美国将这些企业所承诺采取的治理工具推向国际，逐渐转化为全球人工智能治理的通用手段，对全球人工智能治理的具体落地方案发挥了实质性影响。**另一方面，企业探索开发新技术治理策略，并创新制度解决方案。**例如，IBM 推出了 AI FactSheets 360 网站，提供关于人工智能模型重要特征（包括目的、性能、数据集、特征等）的“情况说明书”组装方法。[54]OpenAI 提供“版权盾”（Copyright Shield）侵权赔偿基金，解决用户因使用 GPT 而引发的版权纠纷；提出漏洞赏金计划等。Anthropic 公司发起人工智能宪法（Constitutional AI）公众参与计划，探索公众价值观和不同偏好。[55]

值得注意的是，近期产业竞争加剧，危及企业自愿承诺落地实践。2023 年以来，谷歌、Meta、亚马逊等厂商接连升级大模型产品，激烈竞争局势下 OpenAI 旗下产品用户访问量一度出现负增长。国内来看，大模型厂商打响价格战，各大企业通过降价策略抢占市场份额。例如字节豆包大模型在 B 端市场的定价为 0.0008 元/千 Tokens，比行业便宜 99.3%。**激烈的市场竞争恐迟滞行业的安全治理落地实践。**继 2023 年 11 月 OpenAI“换将风波”之后，2024 年 5 月，OpenAI 超级对齐团队两位负责人接连离职，OpenAI 进而解散了代表公司安全负责任形象的超级对齐团队。微软、谷歌也相继解散道德伦理团队，安

[53] https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administrationsecures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/
[54] IBM Research, ‘IBM Artificial Intelligence Pillars’ (30 August 2023), < https://www.ibm.com/policy/ibm-artificial-intelligence-pillars/> accessed 17 September 2023.
[55] https://www.anthropic.com/news/claudes-constitution

全承诺的效力受到一定影响。

2. 政府引导下专业机构百花齐放，在协同治理中角色凸显

鉴于新一轮人工智能技术的复杂性、不可控性提升，政府支持引导下专业力量的兴起，在优化人工智能技术治理方面发挥了举足轻重的作用。**一是美国国家标准与技术研究所（NIST）为代表的标准化机构。**通过组建生成式 AI 公共工作组[56]，集结了 2500 多名技术专家与志愿者，推动 IBM、Google、OpenAI 等企业内部风险管理流程与 NIST 风险管理框架对接，同时相关成果被纳入拜登政府《14110 号行政令》。**二是以新加坡 AI Verify 基金会为代表的公益开放组织。**在新加坡信息、通信及媒体发展管理局（IMDA）支持下，基金会通过开放编辑和全球协作开发 AI 检测工具，共享人工智能测试框架、代码库、标准和最佳实践。例如，联合其他设计伙伴推出 Project Moonshot，初步弥补早期工具包在测试大语言模型方面的能力缺陷。[57]国际层面，推进 AI Verify 工具包与欧盟、经合组织和新加坡等多个国际监管框架对齐，提升国际合作软实力。**三是以人工智能安全研究所为代表的安全研发机构。**自 2023 年 11 月以来，英、美、日、韩等国依托政府部门或科研机构组建人工智能安全研究所，如，美国设在商务部下属国家标准与技术研究院，日本设在经产省管辖的信息处理推进机构（IPA）等。**国内来看**，包括中国信通院、中国科学院、上海人工智能实验室、北京智源人工智能研究院、清华大学、北京大学等在内的科研机构、高等院校等在人工智能安全基准、技术标准、评

[56] https://www.nist.gov/news-events/news/2023/06/biden-harris-administration-announces-new-nist-public-working-group-ai

[57] https://aiverifyfoundation.sg/project-moonshot/

测工具等方面已具有良好积淀。总体来看，各方安全研究机构在打造人工智能安全标准和指南、推进人工智能安全评估与测试、推进安全研究国际合作等方面均发挥重要作用。

3. 行业社区组织形式日益多样化，构建高效开放的生态圈

企业等私人主体积极搭建产业联盟，探索提升人工智能安全性能的坐标尺。国际层面，2023 年 12 月，Meta、IBM 等全球 50 多个创始成员和合作者，联合发起成立人工智能联盟（AI Alliance）[58]，开发和部署基准和评估标准、工具和其他资源，以加速 AI 技术负责任的创新与发展。2024 年 2 月，包括 200 多家非营利组织、大学、研究团体和企业的美国人工智能安全研究联盟宣布成立[59]，与国际合作伙伴一同制定人工智能模型合作研究和安全指南、开发和推广风险管理工具，并建立人工智能安全事件应急响应机制。**国内层面**，中国人工智能产业发展联盟（AIIA）提出“人工智能风险管理体系”[60]，起草并组织企业签署《人工智能安全承诺》，推出针对大模型、人脸识别等典型应用的风险防控标准，形成安全防御的“坐标尺”，并建设了大模型安全动态评测系统、内容标识平台，建设安全防控的“工具箱”。**此外，通过开源社区建设、赛事激励机制、国际论坛的思想碰撞等形式[61]，逐步形成多元且敏捷的 AI 治理平台。**这一平台不仅为行业提供了丰富的交流与合作机会，还在技术创新与应用推广方面构

[58] https://china.newsroom.ibm.com/2023-12-07-IBM-Meta-50-AI
[59] https://www.commerce.gov/news/press-releases/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated
[60] https://mp.weixin.qq.com/s/wfCAEHY_hryA8Rr9L8MHMQ
[61] 如 2024 年 10 月举办的第五届中国人工智能大赛，https://ai.xm.gov.cn/competition/competition-detail.html?id=2024c40dbb2347fba8b3c9a6294efa5b

建了一个高效开放的生态圈。

来源：中国信息通信研究院

图 9 中国人工智能行业自治图谱

（三）国际人工智能治理日益深化，三大维度热点持续深化

新一轮人工智能技术创新导致治理举措急剧增加，国家、国际和区域组织等竞相发布原则、宣言、声明等，全球人工智能治理格局复杂而分散，包容性有待进一步提升。

1. 机制层面，联合国酝酿新的人工智能治理协调机制

联合国致力于实现全球人工智能治理的协调和统合，新进程建设进入窗口期。2024 年 5 月发布的《联合国系统人工智能治理白皮书》指出，联合国系统在技术、气候、裁军、卫生等领域的规范制定和共识建立、能力发展和国际合作方面发挥着独特作用，展现了其在引导和塑造全球人工智能治理框架中具有重要潜力。2024 年 3 月，联合国大会通过了首个有关人工智能的全球决议《抓住安全、可靠和值得

信赖的人工智能系统带来的机遇，促进可持续发展》[62]，强调人工智能系统伦理、人权保护、数据安全，支持 17 项可持续发展目标。2024 年 7 月，联合国大会通过中国主提的《加强人工智能能力建设国际合作决议》，强调通过国际合作帮助各国尤其是发展中国家加强人工智能能力建设，促进人工智能的包容性、普惠性和可持续发展。9 月 22 日，联合国未来峰会发布《全球数字契约》，就联合国框架下全球人工智能治理达成初步方案。[63]**一方面，联合国内部现有机构积极依托现有框架，推进全球人工智能治理落地。**例如国际电信联盟（ITU）作为联合国负责信息通信技术事务的专门机构，担任联合国系统机构间人工智能工作组联合主席，在人工智能标准制定、监管政策协调等方面具有独特优势，通过建立“人工智能惠及人类”（AI for Good）全球公益平台及峰会、发布《联合国系统人工智能治理白皮书》等方式推动人工智能治理全球合作。**另一方面，酝酿新的协调机制，提出设立发挥粘合剂作用的协调机构。**9 月 19 日，联合国人工智能高级别咨询机构终期报告发布，核心是在联合国秘书处下设立人工智能办公室，以应对当前人工智能有关倡议在代表性、协调性和实施性上面临的碎片化挑战。[64]报告还提出建立人工智能科学小组、政策对话、标准交流中心、能力发展网络和基金、数据框架等七项建议，强调从民用航空、海上作业或核能全球治理的实体模式中吸取与审计和监测程序有关的经验教训。

[62] https://www.state.gov/united-nations-general-assembly-adopts-by-consensus-u-s-led-resolution-on-seizing-the-opportunities-of-safe-secure-and-trustworthy-artificial-intelligence-systems-for-sustainable-development/
[63] https://www.un.org/zh/documents/treaty/A-RES-79-1-Annex-I
[64] https://news.un.org/zh/story/2024/10/1132941

2. 议题层面，人工智能发展优先与风险管控路线并行

一方面，国际社会提出以能力建设为抓手弥合智能鸿沟，推动实现可持续发展目标。联合国在推动全球人工智能能力建设方面发挥关键作用。2024 年 9 月，联合国全球未来峰会聚焦 AI 加剧数字鸿沟的风险，围绕“为人类管理人工智能”主题，探讨数据、技术与治理的互联互通。峰会通过《全球数字契约》，倡导“以人为本的 AI 治理”，鼓励成员国加强公共教育以提升数字技能，并敦促电信运营商增强偏远地区网络覆盖，确保 AI 发展惠及全人类。[65]**多边组织支持发展中国家提升人工智能研发与应用能力。**二十国集团（G20）发布《G20 数字鸿沟倡议》及《数字公共基础设施系统框架》，提供数字基础设施资助，增强数据可及性与安全性。经济合作与发展组织（OECD）通过分享政策与经验，发布 AI 政府间政策指导方针，帮助制定 AI 发展战略。**中国高度关注全球南方的需求，致力于成为能力建设的倡导者与实践者，推动公平普惠的全球治理。**自第 78 届联合国大会通过《加强人工智能能力建设国际合作》决议后，中国迅速采取行动，发布《人工智能能力建设普惠计划》，提出“五大愿景”及“十项行动”，覆盖基建、研发、政策交流等关键领域。[66]2024 年 9 月，中国与联合国秘书处在上海共办首届 AI 能力建设研讨班，近 40 国代表受益。中方还计划至 2025 年底举办 10 期研修项目，并倡议成立国际合作之友小组，持续推动普惠计划，确保全球南方平等共享 AI 发展成果。[67]

另一方面，前沿人工智能安全问题备受国际社会关注。前沿人工

[65] https://www.un.org/zh/summit-of-the-future
[66] https://www.mfa.gov.cn/wjbzhd/202409/t20240926_11497650.shtml
[67] https://www.idcpc.gov.cn/ldt/202412/t20241205_166091.html

智能的风险较一般人工智能体现出更大的不可控性、未知性，加州大学伯克利分校的知名专家斯图尔特·罗素（Stuart Russell）教授指出，人工智能系统可能在新应用场景中暴露出难以预料的灾难性缺陷。**在组织层面，人工智能安全峰会提供全球对话平台，推动建立人工智能安全研究合作网络。**在 2023 年 11 月的英国人工智能安全峰会上，与会各方签署了《布莱切利宣言》，强调共同努力设定共同的监管方法，以应对 AI 带来的机遇和风险。在 2024 年 5 月韩国首尔人工智能安全峰会上，美、英、日等 10 个国家联合发布“全球人工智能安全研究所合作网络”计划，推动全球人工智能安全政策和技术的协同。[68]**2025 年法国安全峰会将进一步推动各方在安全阈值、安全评估评测等方面达成共识。**法国峰会将发布正式版《先进人工智能安全国际科学报告》，对前沿人工智能的迅速发展进行科学评估；将讨论有关前沿人工智能“严重风险”阈值提案；在全球人工智能安全研究网络下，将进一步推进人工智能安全评估评测等方面合作。

3. 执行层面，各方推动高层次承诺向可执行政策落地

国际组织密集推出各类治理举措，成为全球人工智能治理首要供给方。根据欧洲委员会正在编制的全球人工智能监管倡议清单，国际组织在 2020 年超过国家行为体成为此类倡议的主要来源。**二十国集团兼顾全球南北发展需要，巴西峰会聚焦发展问题。**2023 年 9 月，G20 在新德里峰会提出《二十国集团领导人新德里峰会宣言》，推动建立维护全球数字公共基础设施储存库（GDPIR）。2024 年 11 月，

[68] https://www.csis.org/analysis/ai-seoul-summit

G20 在巴西里约热内卢召开会议，充分肯定人工智能在推动持续发展方面的关键作用，强调支持发展中国家的能力建设，推进“促进创新的人工智能治理”，并将于 2025 年设立人工智能工作组，进一步深化国际合作。**七国集团推广人工智能治理共同价值，争取人工智能规则制定主动权。**2023 年 10 月，G7 就承诺执行《广岛进程组织开发先进人工智能系统的国际指导原则》和《广岛进程组织开发先进人工智能系统的国际行为准则》发表联合声明。2024 年 3 月，G7 数字部长提出将广岛人工智能进程框架推广到 G7 成员国之外，试图扩大 G7 集团在国际治理中的影响力。2024 年 5 月，七国集团首脑在广岛召开峰会，同意在广岛 AI 进程框架下成立一个工作组，协调各国 AI 监管事项。[69]**经济合作与发展组织聚焦人工智能发展议程，协助制定人工智能治理政策。**2024 年 5 月，经合组织部长理事会会议通过了《经合组织人工智能原则》的修订文本，强调对人工智能全生命周期商业行为进行治理。[70]**东盟推崇“发展优先”的治理观，为地区人工智能治理提供软性指引。**2024 年 2 月，东盟发布《东盟人工智能治理与伦理指南》，阐述东盟“发展优先”的治理观，提出支持人工智能人才、人工智能创新等国家级建议，以及成立东盟人工智能治理工作组、开发针对人工智能生成物的治理框架等区域级建议。**金砖峰会着力确保全球南方国家共享人工智能红利。**金砖国家扩员后的首届峰会于 2024 年 10 月在俄罗斯举行，峰会推动人工智能技术、政策、人才、产业等方面交流合作，提出人工智能金砖方案。2024 年 7 月，

[69] https://www.csis.org/analysis/shaping-global-ai-governance-enhancements-and-next-steps-g7-hiroshima-ai-process
[70] https://oecd.ai/en/ai-principles

我国成立中国—金砖国家人工智能发展与合作中心，推动金砖国家间在人工智能领域加强信息交流和技术合作、深化产业对接和项目合作、应用赋能和能力提升、治理合作和标准规范建设等。[71]

四、人工智能治理对策建议

人工智能作为我国实施创新驱动发展战略的关键力量，是加快形成新质生产力的重要引擎。展望未来，人工智能技术带来的发展红利和全球性挑战同步显现，各国围绕人工智能评估评测、知识产权、价值对齐等加速搭建规则标准体系，人工智能治理加速迈入实践落地和国际合作的关键窗口。

（一）深化落地人工智能协同敏捷治理模式

探索敏捷治理理念，建立灵活性、全面性制度框架，推动人工智能高质量发展和高水平安全实现良性互动。**一是统筹发展和安全。**通过敏捷治理实现多项目标的平衡，既要摒弃以往以牺牲安全为代价的粗放增长，又要避免严格监管带来的创新抑制效应。支持相关机构在基础创新、数据互通、应用赋能、风险防范等方面开展协作。**二是推动跨部门协同治理。**深入推进跨部门综合监管，是加快转变政府职能、提高政府监管效能的重要举措。应加强跨部门监管能力建设，完善跨部门综合监管协同方式，有力整合央地各方技术支撑力量，深化监管事项清单管理、信息共享、联合检查等机制，跨部门、跨区域、跨层级实施综合监管，避免出现治理真空、治理竞次等问题。**三是建立健全多元敏捷互动机制。**塑造以政府为主导，企业实施自我管理、行业

[71] https://mp.weixin.qq.com/s/1udADFXVAUc8AfruPNF5ig

加强自律规范、社会广泛监督的全方位社会共治模式。政府引导企业查找问题，协调解决试点企业相应困难。企业定期报送风险阶段性评估报告，建立完善内部合规监测与预警机制，在发生重大风险后及时上报事件情况。

（二）系统监测预警人工智能伦理社会影响

为有效应对具身智能、数字人等应用带来的情感依赖、劳动替代等伦理挑战，应对相关伦理影响进行系统监测预警。**一是科学设定监测目标**。选择可追踪、可量化的目标指标进行定期调查，如重点行业的就业替代情况，重点区域的人工智能能源消耗情况等，在建立独立调研渠道的基础上，汇总其他各来源数据，如研究机构、企业、国际组织等的相关报告，确保监测数据来源的客观性、全面性。**二是加强前瞻预警，制定风险预案**。集结经济学、社会学、计算机科学、伦理学、能源科学等领域的专家，构建跨学科团队，确保多角度分析能力。研究划定人工智能在重点领域伦理风险的临界风险。**三是引导公众提升智能素养**。通过电视广播、社交媒体、学校教育等渠道开展广泛的宣传活动，普及 AI 基础知识，探讨人工智能对情感和社会互动的影响，同时强调公众伦理和安全意识，培养健康的情感管理技能。

（三）围绕要素和场景细化负责任创新方案

面对人工智能带来的制度性挑战，针对产业链各环节，坚持“小步快跑”“边发展边治理”的制度路径具有重要现实意义。**一是**围绕算力、模型、数据，医疗、教育等“要素+场景”思路率先制定低位阶立法、标准规范。例如针对数据+医疗、模型+金融等情形细化制度

方案。**二是**破除高质量数据集建设的制度障碍。推进完善数据共享流通规范，推进数据资源互通。推动版权激励方案，探索在著作权法等立法中增加“文本与数据挖掘”例外条款，优化关于个人信息使用的合法性基础，健全具体场景下个人信息保护实施细则。**三是以监管沙箱试点探索成熟的分级分类方案。**建议推进应用成熟、制度需求大的行业领域，开展人工智能治理试点工作，鼓励企业积极试行治理方案和工具，摸清主要场景和关键环节的风险问题，明确细化分级分类的可操作标准，并根据实际情况做动态调整。

（四）立足于全生命周期优化安全技术工具

人工智能治理既需要完善治理理念与规则，也需要优化治理手段与能力，进一步更新丰富治理工具箱。**一是探索安全技术研发创新。工具，推进事前、事中、事后全流程监管。**广泛发掘一批可推广复制的技术路线，不断丰富内容标识、红队测试、价值对齐等技术工具和治理经验，发布数据审查库、数据标注规范等具体评估指引，增强人工智能风险的动态感知、科学预警、留痕溯源能力。**二是强化人工智能监管平台资源配备。**建立国家级安全基准测试平台，促进产学研用各创新主体联动和共建共享，通过政策扶持、竞赛奖励、创新支持等方式，推进评估评测方法、测试数据集、可信数据库建设，推进发布规则规范、测试框架和标准体系。**三是加强人工智能领域第三方评估机构力量。**明确人员专业能力、技术工具储备、资源平台建设等资质认定条件，并定期进行资质年审。鼓励科研院所、协会、企业等挖掘一批专业过硬的人才重点培养，建设专家力量储备，就重大议题加强

研究储备，共同构建优势互补、协同发展的安全研究网络。

（五）加速落地全球人工智能能力建设方案

人工智能治理攸关全人类命运，是世界各国面临的共同课题，应积极参与并推动国际合作治理。**一是推动建立人工智能能力建设平台。**落实联合国未来峰会成果，深化人工智能产供链国际合作，开展基础设施建设、研发和应用合作，协助落后国家人才培养和教育培训，促进人工智能技术公平获取和安全使用。**二是协助后发国家加快人工智能治理能力建设。**推动安全技术开源开放共享，积极分享在人工智能测试、评估、认证与监管方面的政策与技术实践。通过国际标准化组织（ISO）、国际电工委员会（IEC）和国际电信联盟（ITU）等平台，围绕关键术语等开展标准研究，共同探索国际互认的测试评估方法。**三是推进公平普惠发展。**推动人工智能数据语料库平等多样，消除种族主义、歧视和其他形式的算法偏见，促进、保护和保全语言和文明多样性。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62305772

传真：010-62302476

网址：www.caict.ac.cn