CAICT 中国信通院

No.202205

人工智能白皮书

（2022 年）

中国信息通信研究院

2022年4月

版权声明

前　言

在新科技革命和产业变革的大背景下，人工智能与产业深度融合，是释放数字化叠加倍增效应、加快战略新兴产业发展、构筑综合竞争优势的必然选择。当前，人工智能加快向各产业渗透，正在促进新兴产业之间、新兴产业与传统产业之间以及技术与社会的跨界融合发展。在“十四五”开端，全面梳理人工智能的发展态势，具有十分重要的参考意义。

本白皮书重点从人工智能政策、技术、应用和治理等维度进行了阐述。**政策层面**，国内外不断强化人工智能的战略地位，推动释放人工智能红利。**技术及应用层面**，以深度学习为代表的人工智能技术飞速发展，新技术开始探索落地应用；工程化能力不断增强，在医疗、制造、自动驾驶等领域的应用持续深入；可信人工智能技术引起社会广泛关注。与此同时，**治理层面**工作也受到全球高度关注，各国规制进程不断加速，基于可信人工智能的产业实践不断深入。

总体来看，本白皮书认为人工智能逐步进入新的阶段，下一步的发展方向，将由**技术创新、工程实践、可信安全**“三维”坐标来定义和牵引。具体来看，第一个维度突出创新，围绕着算法和算力方面的创新仍会不断涌现。第二个维度突出工程，工程化能力逐渐成为人工智能大规模赋能千行百业的关键要素。第三个维度突出可信，发展负责任和可信的人工智能成为共识，将抽象的治理原则落

实到人工智能全生命流程将成为重点。

由于人工智能发展速度之快、辐射范围之广、影响程度之深前所未有，我们对人工智能的认识还有待进一步深化，白皮书中存在的不足之处，欢迎大家批评指正。

目 录

图 目 录

一、人工智能发展概述

人工智能是引领未来的新兴战略性技术，是驱动新一轮科技革命和产业变革的重要力量。习近平总书记多次作出重要指示，强调“要深入把握新一代人工智能发展的特点，加强人工智能和产业发展融合，为高质量发展提供新动能”。近年来，人工智能相关技术持续演进，产业化和商业化进程不断提速，正在加快与千行百业深度融合。站在“十四五”开端这一特殊的节点，我们坚信全面梳理人工智能政策、技术、应用以及治理的发展态势，能够有益于凝聚业界共识，推动人工智能持续健康发展。

（一）全球不断升级人工智能战略，纷纷抢抓重要发展机遇

人工智能已成为科技创新的关键领域和数字经济时代的重要支柱。自 2016 年起，先后有 40 余个国家和地区将推动人工智能发展上升到国家战略高度。近两年来，特别是新冠疫情的冲击下，越来越多的国家认识到，人工智能对于提升全球竞争力具有关键作用，纷纷深化人工智能战略。**欧盟**发布《2030 数字化指南：欧洲数字十年》、《升级 2020 新工业战略》等，拟全面重塑数字时代全球影响力，其中将推动人工智能发展列为重要的工作。**美国**陆续成立了国家人工智能倡议办公室、国家 AI 研究资源工作组等机构，各部门密集出台了系列政策，将人工智能提到“未来产业”和“未来技术”

领域的高度，不断巩固和提升美国在人工智能领域的全球竞争力，确保“领头羊”地位。**日本**继制定《科学技术创新综合战略 2020》之后，于 2021 年 6 月发布“AI 战略 2021”[1]，致力于推动人工智能领域的创新创造计划，全面建设数字化政府。**英国**于 2021 年 9 月发布国家级人工智能新十年战略，这是继 2016 年后推出的又一重要战略，旨在重塑人工智能领域的影响力。**中国**《中共中央关于制定国民经济和社会发展第十四个五年规划和 2035 远景目标纲要的建议》指出，要瞄准人工智能等前沿领域，实施一批具有前瞻性、战略性重大科技项目，推动数字经济健康发展。

面向人工智能领域创新需求的投资不断加大。主要经济体通过激励计划和直接投资项目等推动人工智能发展已广泛实践。**欧盟不断加大人工智能产业资金支持力度，大力促进欧洲的数字变革。**欧盟有史以来最大的支持研发和创新项目——“地平线欧洲”计划总投资额达 955 亿欧元，明确将人工智能列入资金支持范围。2021 年 4 月，欧盟以条例的形式通过“数字欧洲计划”，对包括人工智能在内的项目进行投资，总额达 75.9 亿欧元[2]。**美国以保持领先地位为战略目标并持续加大人工智能领域投入。**美国 2021 年人工智能非国防预算增加约 30%，总额达到 15 亿美元[3]。此外在《美国创新与竞争法案》中，将人工智能、量子计算等列为 2022 财年美国研发预算优

[1] https://www8.cao.go.jp/cstp/ai/aistrategy2021_gaiyo.pdf

[2] 资料来源：欧盟委员会。

[3] https://www.sohu.com/a/413600243_115978

先事项，未来对包括人工智能在内的多个领域共投入 1000 亿美金进行研发工作。**英国将投资和规划人工智能生态系统作为长期战略**，启动国家人工智能研究与创新计划，支持人工智能先进研究等。据统计，2014 年到 2021 年之间对人工智能的投资已经超过 23 亿英镑[4]。

通过应用牵引推动人工智能技术落地成为各国共识。美国引导人工智能技术在行业领域的创新和融合应用。2021 年 7 月，美国国家科学基金会联合多个部门和知名企业等，新成立 11 个国家人工智能研究机构，涵盖了人机交互、人工智能优化、动态系统、增强学习等方向，研究项目更是涵盖了建筑、医疗、生物、地质、电气、教育、能源等多个领域。**英国支持人工智能产业化**，启动人工智能办公室和英国研究与创新局联合计划等，确保人工智能惠及所有行业和地区，促进人工智能的广泛应用。**日本将基础设施建设和人工智能应用作为重点**，提出加快建设相关基础设施，重点强调了跨行业的数据传输平台以及人工智能相关标准等，全面推动人工智能在医疗、农业、交通物流、智慧城市、制造业等各个行业开展应用，并加大对中小企业的支援。**我国十四五规划纲要明确大力发展人工智能产业，打造人工智能产业集群以及深入赋能传统行业成为重点**。2021 年 4 月，工信部支持创建北京、天津（滨海新区）、杭州、广州、成都等第二批国家人工智能创新应用先导区，不断强化应用牵

[4] https://www.gov.uk/government/news/new-ten-year-plan-to-make-britain-a-global-ai-superpower

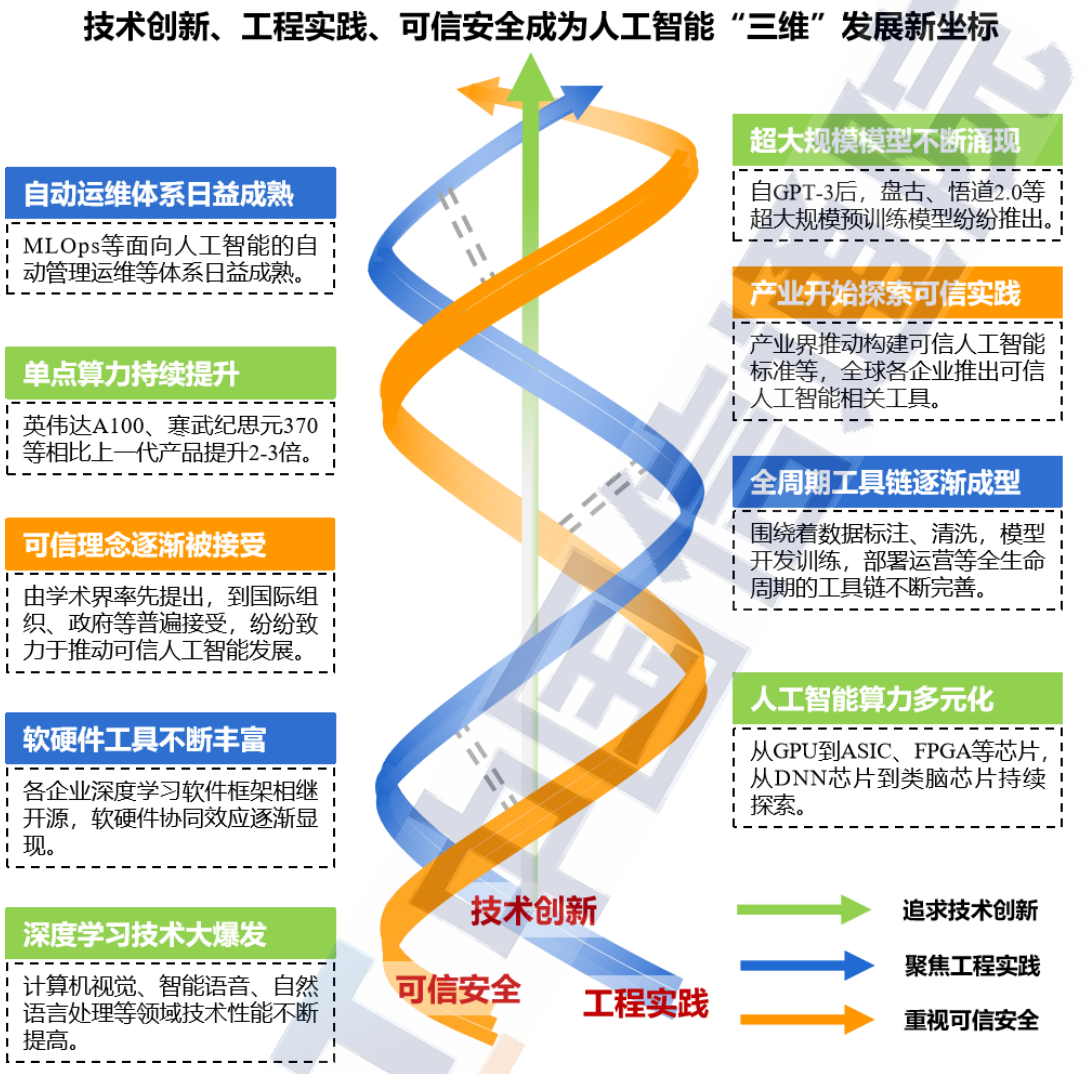
引作用。科技部支持建设多个人工智能创新发展试验区，陆续批复北京、上海、天津、深圳、杭州等 15 个国家新一代人工智能创新发展试验区。

（二）人工智能开始迈入全新阶段，持续健康发展成为焦点

人工智能自 1956 年诞生以来，相关理论和技术持续演进。直到近十年，得益于深度学习等算法的突破、算力的不断提升以及海量数据的持续积累，人工智能才得以真正大范围地从实验室研究走向产业实践。产业发展和赋能的过程中，一方面，大量的实践场景均能看到从“可用”到“好用”的发展路径，这离不开技术自身的持续迭代，工程实现的不断优化，以及管理体系的支撑保障。另一方面，随着人工智能应用暴露出各种风险和挑战，以及人们对人工智能认识的不断深入，人工智能治理已经成为全球各界高度关注的议题，对可信安全的呼声不断增多。

未来人工智能除了重视技术创新以外，还更加关注工程实践和可信安全，这也构成了新的“三维”发展坐标，牵引人工智能技术产业迈向新的阶段。事实上，业界在各个维度上的努力早已开始，并且从未停止过，只是时至今日，工程实践和可信安全被摆在了更为重要的位置。三维坐标并非完全独立，而是相互交织、相互支撑。图 1 给出了本轮人工智能浪潮以来沿着各个方向演进的示意图，概

述了各坐标下的发展脉络。



来源：中国信息通信研究院

图 1 人工智能演进的三个维度示意图

追求特定场景下的技术创新一直是人工智能发展的目标和驱动力。以深度学习为代表的算法爆发拉开了人工智能浪潮的序幕，在计算机视觉、智能语音、自然语言处理等领域广泛应用，相继超过

人类识别水平[5]。人工智能算力的多元化以及单点算力的不断提升，有力支撑了人工智能的发展。再到近期，国内外超大规模预训练模型频繁涌现，不断刷新各个应用领域的榜单。未来，在算法、算力等方面仍将持续变革，为迈向更加智能的时代奠定基础。

工程实践能力日益成为释放人工智能技术红利的重要支撑。在工程实践方面的努力，最早可追溯至 Caffe、TensorFlow、PaddlePaddle 等开源框架的诞生，通过屏蔽底层硬件和操作系统细节，大幅降低模型开发和部署难度，有效推动了人工智能技术的扩散。当前，人工智能与云计算、大数据等支撑技术的融合不断深入，围绕着数据处理、模型训练、部署运营和安全监测等各环节的工具链不断丰富。人工智能研发管理体系日益完善，以 MLOps 为代表的自动运维技术受到越来越多的关注。随着工程实践能力的不断提升，“小作坊、项目制”的赋能方式正在成为历史，未来将会更加便捷、高效地实现人工智能落地应用和产品交付。

可信安全逐渐成为人工智能赋能过程中不可或缺的保障。可信人工智能最早由学术界提出，近年来围绕着安全性、稳定性、可解释性、隐私保护、公平性等方面的可信人工智能研究持续升温。可信人工智能理念得到了国际组织的广泛关注，二十国集团（G20）在 2019 年 6 月提出的“G20 人工智能原则”中明确建议促进可信赖的人

[5]https://ai100.stanford.edu/2021-report/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence

工智能创新发展，成为了重要的共识。可信人工智能的理念逐步贯彻到人工智能的全生命周期之中，产业实践不断丰富，已经演变为落实人工智能治理相关要求的重要方法论。

总的来看，人工智能正在迈入**“创新驱动、应用深化、规范发展”的新阶段**。从人工智能自身产业化的角度来看，技术迭代升级是发展的源动力，目前人工智能尚不完善，智能化路径还在加快探索，技术的创新驱动将有助于拓展新的发展空间。从人工智能赋能传统产业的角度来看，特别是疫情以来，数字化、智能化转型不断提速，推动人工智能应用迈入加速轨道，相关应用不断深化。从治理角度来看，技术和产业发展要领先于监管和制度，治理问题日益严峻，保障人工智能的健康发展成为全球共同关注。这里面既有渐进的变化，也有结构性甚至方向性的调整，需要全面、系统地提升各方面能力，从而推动人工智能持续且健康的发展。

二、人工智能技术及应用沿着“创新、工程、可信”三个方向持续演进

在新的背景下，人工智能技术亦需要适应新的变化。本章按照新三维坐标对人工智能技术及应用的发展态势进行了梳理。围绕着算法、算力和数据的技术创新始终是前进主旋律；工程实践中的相关技术开始覆盖人工智能全流程，加速人工智能大规模落地应用；人工智能可信技术是破解治理难题的重要支撑，愈发受到各界关注。

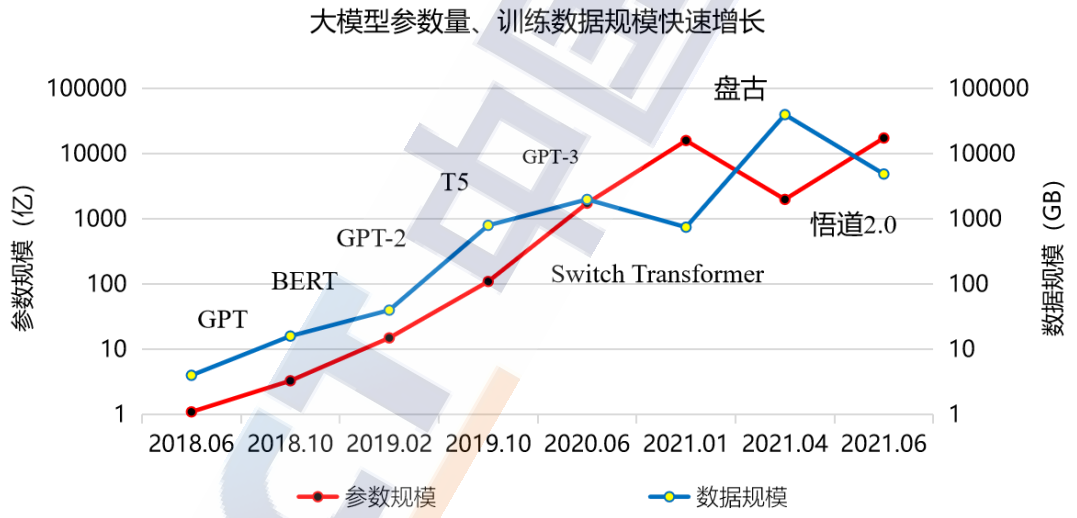
（一）人工智能在追求极致创新方面不断突破

一直以来，算法、算力和数据被认为是人工智能发展的三驾马车，也是推动人工智能发展的重要基础。**在算法层面**，超大规模预训练模型等成为近两年最受关注的热点之一，不断刷新各领域榜单[1, 2]；知识驱动的人工智能等方向研究成为提升认知能力的重要探索[3, 4]；人工智能与各科学研究领域的融合创新日益受到关注，人工智能成为基础科学研究的重要工具。**在基础算力层面**，单点算力持续提升，算力定制化、多元化成为重要发展趋势；计算技术围绕数据处理、数据存储、数据交互三大能力要素演进升级，类脑芯片、量子计算等方向持续探索[5]。**在数据层面**，以深度学习为代表的人工智能技术需要大量的标注数据，这也催生了专门的技术乃至服务，随着面向问题的不断具体化和深入，数据服务走向精细化和定制化；此外，随着知识在人工智能的重要性被广泛提及，对知识集的构建和利用不断增多。

1. 新算法不断涌现，技术融合成重要趋势

超大规模预训练模型推动技术效果不断提升，继续朝着规模更大、模态更多的方向发展。自 OpenAI 于 2020 年推出 GPT-3 以来，谷歌、华为、智源研究院、中科院、阿里巴巴等企业和研究机构相继推出超大规模预训练模型，包括 Switch Transformer、DALL · E、MT-NLG、盘古、悟道 2.0、紫东太初和 M6 等，不断刷新着各榜单

纪录，百度 ERNIE3.0 模型[6]在自然语言理解任务上的综合评分（GLUE）已达 90%以上，智源悟道文澜模型[7]在多源图文数据集评分（RUC-CAS-wenlan）相比 OpenAI 的 CLIP 模型大幅提升 37.0%。当前，预训练模型参数数量、训练数据规模按照 300 倍/年的趋势增长，继续通过增大模型和增加训练数据仍是短期内演进方向；另外，跨模态预训练大模型日益普遍，已经从早期只学习文本数据，到联合学习文本和图像，再到如今可以处理文本、图像、语音三种模态数据，未来使用更多种图像编码、更多种语言、以及更多类型数据的预训练模型将会涌现，这也是实现人工智能通用化的有益探索。



来源：中国信息通信研究院

图 2 大模型参数量和训练数据规模增长示意图

轻量化深度学习技术不断探索，计算效率显著提升。复杂的深度学习模型往往需要消耗大量的存储空间和计算资源，难以在端、边等资源受限情形下应用，具备低内存和低计算量优势的技术成为

业界需求。轻量化深度学习成为解决这一挑战的重要技术，包括设计更加紧凑和高效的神经网络结构、对大模型进行剪枝（即“裁剪”掉部分模型结构），以及对网络参数进行量化从而减少计算量等方向。例如，谷歌提出的 MobileNet 和旷视提出 ShuffleNet 等成为紧凑模型的典型代表；百度推出的轻量化 PaddleOCR 模型规模减小至 2.8Mb，在 GitHub 上开源后受到热捧[6]。

“生成式人工智能”技术不断成熟，未来听、说、读、写等能力将有机结合起来。目前，“生成式人工智能”技术被广泛应用于智能写作、代码生成、有声阅读、新闻播报、语音导航、影像修复等领域，通过机器自动合成文本、语音、图像、视频等正在推动互联网数字内容生产的变革。听、说、读、写等能力的有机结合成为未来发展趋势。例如央视、新华社、光明网等均推出了数字人主播，支持从音频/文本内容一键生成视频，能够实现节目内容快速、自动化生产，相关数字人主播和数字人记者，已在全国两会、春节晚会等大型报道和节目中广泛应用。

知识计算成为推动人工智能从感知智能向认知智能转变的重要探索。知识凝聚了人的智慧，知识和数据的双驱动有助于解决不完全信息、不确定性和动态环境下的推理决策问题，可以提高人工智能系统的智能化水平。目前，围绕着知识获取、知识建模、知识管

[6] https://github.com/PaddlePaddle/PaddleOCR

理、知识应用等过程，已经形成了涵盖知识图谱、知识库、图计算等技术，覆盖知识表示、知识计算、知识推理与决策能力的体系，可实现对知识的管理与利用。学术界和产业界都已经开始推出基于知识的人工智能应用平台或解决方案，例如清华大学、浙江大学、华为云、智源研究院、百度、竹间智能、国双等推出的知识计算引擎、知识中台、知识工程平台、知识智能平台等解决方案。未来，知识计算将着力在深度学习算法中嵌入先验知识建立可解释模型，让知识深入参与模型求解，进一步提高人工智能的效率、水平以及鲁棒性、可解释性、可迁移性。

人工智能与科学研究融合不断深入，开始“颠覆”传统研究范式。近年来，人工智能对海量数据的分析能力能够让研究者不再局限于常规的“推导定理式”研究，可以基于高维数据发现相关信息继而加速研究进程。2020 年，DeepMind 提出 AlphaFold2 在国际蛋白质结构预测竞赛（CASP）拔得头筹，能够精确地预测蛋白质的 3D 结构，其准确性可以与使用冷冻电子显微镜等实验技术解析的 3D 结构相媲美。中美研究团队使用 AI 的方法，在保证“从头计算（ab initio）”高精度的同时，将分子动力学极限提升了数个量级，比过去同类工作计算空间尺度增大 100 倍，计算速度提高 1000 倍，获得 2020 年 ACM 戈登贝尔奖[7]。更为惊喜的是人工智能与力学、化学、材料学、

[7] https://www.sciencenet.cn/skhtmlnews/2021/3/4443.html

生物学乃至工程领域等融合探索不断涌现，未来将不断拓展人工智能应用的深度和广度。

2. 单点算力持续突破，新技术仍处于探索阶段

当前人工智能算力持续突破，面向训练用和推断用的芯片仍在快速演进。这主要源于算力需求的驱动，一方面体现在模型训练阶段，根据 Open AI 数据，模型计算量增长速度远超人工智能硬件算力增长速度，存在万倍差距[8]；另一方面，由于推断的泛在性，使得推断用算力需求持续增长。与此同时，新的算力架构也在不断研究中，类脑芯片、存内计算、量子计算等备受关注，但总体上处于探索阶段。

训练芯片创新加速，推断芯片朝着专用定制化发展。基于 GPU 的训练芯片持续增多，面向 GPU 创新的企业开始发力，出现了摩尔线程、天数智芯、壁仞科技等一批专注 GPU 赛道的初创公司。**基于 ASIC 等架构云端训练芯片能力提升显著，**寒武纪的思元 370、燧原科技的“邃思 2.0”以及百度的昆仑 2 等相对上一代产品均有 3-4 倍以上的算力提升。**专用定制的端侧推理芯片百花齐放，面向手机应用的智能芯片成为亮点。**2021 年 1 月，联发科推出了高端手机芯片 Dimensity 1200，可边缘处理 5G、AI 和图像数据等。8 月，谷歌为其 Pixel 系列手机专门推出了首款智能手机芯片 Tensor。

类脑芯片、存内计算、量子计算等依旧是重点探索方向。类脑

芯片、存内计算、量子计算等技术在理论层面可实现高算力、低功耗等优点，取得了一些进展，但总体上来看目前技术成熟度相对较低。北京大学类脑智能芯片中心在 2021 年 ISSCC 发布“超低功耗智能物联网芯片(AIoT)”等成果。新型人工智能芯片受到投资资金青睐，2021 年以来多家企业完成了亿元级 A 轮或 A+轮融资，包括 3D 视觉 AI 芯片厂商埃瓦科技，专注神经拟态感存算一体芯片研发的九天睿芯，以及 AI 视觉芯片研发公司爱芯科技等。

3. 数据规模不断提升，构建领域知识集成热点

人工智能的快速发展推动数据规模不断提升。据 IDC 测算，2025 年全球数据规模将达到 163ZB，其中 80%-90%是非结构化数据[8]。**数据服务进入深度定制化的阶段，**百度、阿里巴巴、京东等公司推出根据不同场景和需求进行数据定制的服务；企业需求的数据集从通用简单场景向个性化复杂场景过渡，例如语音识别数据集从普通话向小语种、方言等场景发展，智能对话数据集从简答问答、控制等场景向应用场景、业务问答等方向发展。

各方积极探索建立高质量知识集，支撑未来知识驱动的人工智能应用发展。知识集中包含语音、图像、文本等传统数据和定义、规则、逻辑关系等，是知识的数据化呈现，业界著名知识集有 Wordnet、Hownet 等。例如阿里巴巴联合香港理工大学基于服装设计知识开发

[8] 《数据时代 2025》 IDC

FashionAI 知识集，加速了 AI 在服装设计产业落地应用。

（二）人工智能工具链成为工程实践能力核心

随着人工智能技术不断发展，近年来工程落地应用呈现加速态势。金融领域，人工智能技术开始深入前台、中台、后台全过程；医疗人工智能开始迈入市场化阶段，截至 2021 年 8 月底，共有 28 款产品获批三类医疗器械注册证；制造领域人工智能快速发展，德勤预计我国未来五年将保持年均 40%以上的增长率。目前企业应用人工智能呈现出从初步探索到规模应用的过渡，总体上来看，不断提升工程实践能力成为未来应用的关键。

人工智能工程化开始成为各界关注焦点。学术界，卡耐基梅隆大学软件工程学研究所于近年启动人工智能工程化研究，并联合高校和工业界承担了一项由美国官方机构资助的国家研究计划；世界知名人工智能专家乔丹（Michael I.Jordan）、邢波等认为人工智能工程化是一门新兴的工程科学，是人工智能从理论学科到工程学科发展的趋势。**产业界，**Gartner 连续两年把人工智能工程化列为年度战略技术趋势之一，阿里云等企业把人工智能工程化视作将 AI 变为企业生产力的关键。

人工智能工程化聚焦工具体系、开发流程、模型管理全生命流程的高效耦合。工具体系层面，体系化与开放化成为研发平台技术工具链的发展特点。围绕机器学习和深度学习等技术，已初步构建

起较为完备的工具体系，大幅降低数据处理、模型开发和部署、运维管理等难度，其中关键的软件框架多采用 TensorFlow、PyTorch、Paddle、MindSpore、OneFlow 等开源框架；**开发流程层面**，工程化关注人工智能模型开发的生命流程，追求高效且标准化的持续生产、持续交付和持续部署，最终以最佳的模型进入应用层面产生商业价值。例如 MLOps 就是为了连接模型构建团队、业务团队和运维团队，建立起标准化的模型开发、部署与运维流程。**模型管理层面**，随着企业智能化应用的逐步加深，模型种类和数量大幅增长，企业需要建设对模型生命周期的管理机制，对模型的版本历程、性能表现、属性、相关数据、衍生的模型档案等进行标准化的管理运维。

自动机器学习技术是提升工程化能力的重要能力。自动机器学习是指在机器学习开发应用全流程的部分环节或者全部环节实现自动化，可以有效降低当前阶段人工智能开发门槛高、技术人才匮乏等挑战。该技术主要包括自动数据预处理、自动特征工程、自动超参数搜索、自动模型网络结构设计、自动模型部署等内容，低代码开发、预训练模型等技术也与自动机器学习密切相关，并呈现融合发展的趋势。当前，头部互联网企业和创新企业已经开始积极布局 AutoML 技术和工具，但受限于技术成熟度，AutoML 的应用场景还停留在某些开发环节（如特征工程）或者某些特定的技术领域（如语音识别、目标检测、智能对话等）。

云边端协同管理的技术需求逐渐凸显，人工智能上云进程不断

加速。随着人工智能与各个行业的深度融合，人工智能边缘端和终端设备将得到越来越广泛的应用，同时开发者也将面临边端设备繁杂不易适配、运维管理难等问题。一方面，平台通过模型压缩、自适应模型生成等技术实现边端设备的模型适配和部署；另一方面，通过对编译优化、中间表示等的设计和配置，实现云边端设备的协同管理和运维。

（三）安全可信人工智能技术朝着一体化发展

随着社会各界对人工智能信任问题的不断关注，安全可信的人工智能技术已成为研究热点。研究的焦点主要是提升**人工智能系统稳定性、可解释性、隐私保护、公平性等**，这些技术构成了可信人工智能的基础支撑能力[9]。

人工智能系统稳定性技术重点逐步从数字域扩展到物理域。人工智能系统面临中毒攻击、对抗攻击、后门攻击等特有攻击，这加大了安全性方面的挑战。这些攻击技术既可互相独立也可以同时存在，例如通过打印对抗样本眼镜等能够直接对人脸识别系统造成物理层面的干扰，攻击者在路牌上粘贴对抗样本扰动图案，使得自动驾驶系统错误地将“停止”路牌识别为“限速”路牌等[10]。围绕着人工智能系统的稳定性测试技术也成为了关键，华为、百度等纷纷推出基于模糊理论的相关测试技术，致力于探索提高人工智能系统的稳定性。

人工智能可解释性增强技术仍处在初期阶段，多种路径持续探索。增强人工智能系统的可解释性成为热点工作[11, 12, 13]，主要路径包括建立适当的可视化机制尝试评估和解释模型的中间状态；通过影响函数来分析训练数据对于最终收敛的人工智能模型的影响；通过方法分析人工智能模型利用哪些数据特征做出预测；通过使用简单的可解释模型对复杂的黑盒模型进行局部近似来研究黑盒模型的可解释性等。

隐私计算技术助力人工智能数据安全可信地进行协作。人工智能系统需要依赖大量数据，然而数据的流转过程以及人工智能模型本身都有可能泄漏敏感隐私数据。AI 结合隐私计算技术，可从数据源端确保原始数据真实可信。利用隐私计算技术，数据“可用不可见”，形成物理分散的多元数据的逻辑集中视图，可以保证 AI 模型有充足的、可信的数据可供利用。

提升人工智能公平性的关键在于从数据和技术两方面入手。随着人工智能系统的广泛应用，不公平决策行为以及对部分群体的歧视等问题越来越突出，导致这些决策偏见主要原因如下：受数据采集条件限制，不同群体在数据中所占权重不均衡；在不平衡数据集上训练得到的人工智能模型，造成模型决策不公平。为了保障人工智能系统的决策公平性，从数据层面来看，主要通过构建完整异构数据集，将数据固有歧视和偏见最小化；对数据集进行周期性检查，

保证数据高质量性。从技术层面来看，需要通过引入公平决策量化指标的算法，来减轻或消除决策偏差及潜在的歧视。

体系化推进人工智能可信安全技术将是重要趋势。一方面，当前相关研究多是从稳定、隐私、公平等单一维度展开。已有研究工作表明，稳定性、公平性、可解释性等不同要求之间存在相互协同或相互制约的关系，若仅考虑某一个方面的要求则可能会造成其他要求的冲突。如何构建系统的研究框架，从而保持不同特征要素之间的最优动态平衡成为关键。另一方面，需要从系统层面开展可信安全的研究[14]，这一问题不仅仅是人工智能算法层面问题，还涉及到整个系统，例如人工智能承载操作系统、软件框架、第三方库，以及硬件设备自身的安全问题等，需要构建人工智能全链条、全生命周期的可信安全。

三、全球高度关注人工智能治理工作，人工智能安全可信成重点

人工智能发展的空间越大、影响越深、挑战越多，对它的治理就越重要、越紧迫。当前，全球已经形成多元主体参与、协同共治的治理模式，各国及各组织推出了一系列治理原则，立法进程取得实质性进展，行业组织及企业主体积极探索可信落地实践。

（一）人工智能风险不断增多，全球初步建立治理机制

1. 人工智能深入赋能引发挑战

人工智能带来的风险与挑战是多方面的。除了人工智能技术自身存在天然的缺陷外，区别于纯粹的技术风险，人工智能风险的渊源是人工智能系统的应用对现有的规范体系以及伦理与社会秩序的冲击[15]。

人工智能固有技术风险持续放大。以深度学习为核心的人工智能技术正在不断暴露出由其自身特性引发的风险隐患。一是深度学习模型存在脆弱和易受攻击的缺陷，使得人工智能系统的可靠性难以得到足够的信任。二是黑箱模型具备高度复杂性和不确定性，算法不透明容易引发不确定性风险。三是人工智能算法产生的结果过度依赖训练数据，如果训练数据中存在偏见歧视，会导致不公平的智能决策产生。

现有法律及规范体系受到的挑战不断扩大。人工智能对就法律及规范体系造成了多个方面冲击：**在主体资格界定方面**，沙特阿拉伯授予机器人索菲亚以公民资格引发全球争议，此外还产生了人工智能是否能够成为专利的发明者等问题，如 2021 年 7 月澳大利亚联邦法院裁定人工智能系统可被列为专利申请中的发明人，与美国、英国持截然不同的态度[9]。**在隐私保护方面**，人工智能的发展伴随侵犯个人隐私问题时有发生，央视“3·15”晚会曝光，大量企业违规采集

[9]https://www.abc.net.au/news/2021-08-01/historic-decision-allows-ai-to-be-recognised-as-an-inventor/100339264?utm_campaign=news-article-share-2-control&utm_content=twitter&utm_medium=content_shared&utm_source=abc_news_web

顾客人脸信息用于商业目的[10]。**在责任划分方面**，2015 年英国首例机器人手术致人死亡[11]，特斯拉“失控门”事件使得自动驾驶辅助系统受到质疑[12]。

伦理及社会秩序受到的冲击愈发严重。人工智能存在对人类权利造成冲击的风险，人工智能引发歧视、对人类行为提出新规则、劳动力的变革更替等问题。2021 年 8 月，俄罗斯在线支付服务公司 Xsolla 使用算法判断员工“不敬业且效率低”，并解雇了公司占总人数三分之一的 147 名员工[13]。人工智能直接或间接伤害人类，冲击社会秩序。2020 年 11 月有媒体报道伊朗核科学家被“人工智能”控制的武器刺杀[14]，2019 年亚马逊智能音箱曾给出劝人类自杀的建议[15]。

2. 全球掀起人工智能治理浪潮

当前，面临人工智能深入赋能而引发的多方面风险及挑战，全球各国越来越重视人工智能治理。人工智能治理是一项复杂的系统工程，根据《人工智能治理白皮书》[16]，人工智能治理体系由政府、行业组织、企业以及公众等多元主体共同参与、协同合作，形成了

[10] https://www.163.com/tech/article/G55HQCAA00097U7R.html

[11] https://www.guancha.cn/internation/2018_11_08_478891.shtml

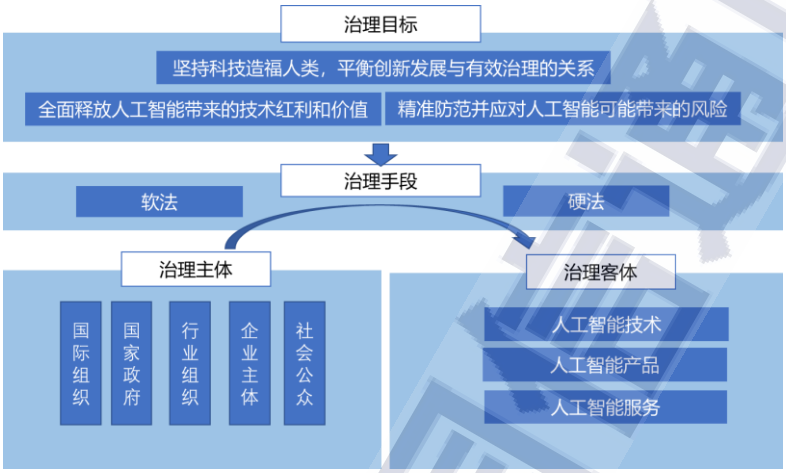
[12] https://new.qq.com/rain/a/20210210A0CPO600

[13]https://gameworldobserver.com/2021/08/04/xsolla-fires-150-employees-using-big-data-and-ai-analysis-ceos-letter-causes-controversy

[14] https://baijiahao.baidu.com/s?id=1685477846545190363&wfr=spider&for=pc

[15] https://www.dailymail.co.uk/news/article-7809269/Amazon-Alexa-told-terrified-mother-29-stab-heart-greater-good.html

伦理原则等“软法”以及法律法规等“硬法”相结合的治理手段，旨在实现科技向善、造福人类的总体目标愿景，推动人工智能健康有序发展。人工智能治理机制如图 3 所示。



资料整理：中国信息通信研究院

图 3 人工智能治理机制示意图

世界主要经济体聚焦人工智能治理焦点议题的讨论，政府间国际组织成为重要发声阵地，联合国、G20、OECD、G7 等成为引导全球人工智能治理的重要组织。OECD 相关研究成果对于推动全球人工智能治理起到了重要作用，是 G7 人工智能监管相关决议、G20 人工智能原则的重要参考。

联合国（UN）积极推动人工智能伦理治理进程。联合国科教文组织（UNESCO）于 2021 年 11 月 25 日发布《人工智能伦理问题建议书》，这是全球首个针对人工智能伦理制定的规范框架，是迄今为止全世界在政府层面达成的最广泛的共识，同时赋予各国在相应层面应用该框架的责任。世界卫生组织（WHO）于 2021 年 6 月 28

日发布了第一份关于在医疗卫生中使用人工智能的指南《医疗卫生中人工智能的伦理和管制》，确保人工智能技术能够为全球所有国家的公共利益服务。

二十国集团（G20）于 2019 年 6 月在参考《OECD 人工智能原则》的基础上，批准了倡导人工智能使用和研发“尊重法律原则、人权和民主价值观”的《G20 人工智能原则》，成为人工智能治理方面的首个政府间国际共识，确立了以人为本的发展理念。中国支持围绕人工智能加强对话，落实 G20 人工智能原则，推动全球人工智能健康发展。

经济合作与发展组织（OECD）于 2019 年 5 月 22 日发布了全球首个人工智能政府间政策指导方针，形成《OECD 人工智能原则》，确立了负责任地管理可信赖的人工智能的五项原则。OECD 于 2020 年 2 月设立了人工智能政策观察站（OECD.AI），通过分享人工智能政策及实践的最佳案例，促进国际合作，帮助成员国打造可信赖的人工智能系统以实现全社会利益。

七国集团（G7）在全球发达经济体间就人工智能治理开启了共识性的探索。2021 年 1 月举行的 G7 峰会表示各成员国将合作研究国际人工智能标准[16]；9 月，在 G7 数据保护和隐私当局会议上，表示未来将把数据保护和隐私监管作为人工智能治理的核心工作，并

[16] https://www.digitalhealth.net/2021/01/hancock-uk-will-work-with-g7-and-others-to-look-at-ai-standards/

推动业界设计满足数据保护要求的人工智能产品。

(二)人工智能治理迈入软硬法协同和场景规制新阶段

自 2017 年《阿西洛马人工智能原则》问世以来，全球掀起了探索制定人工智能伦理原则的热潮。当前，G20 人工智能原则被国际社会普遍认同，政府间组织成为引导人工智能治理方向的重要力量，全球各国加速完善人工智能治理相关规则体系。2021 年欧盟率先推出《人工智能法》草案标志着人工智能治理从原则性约束的“软法”向更具实质性监管的“硬法”加速推进；与此同时，随着人工智能与实体经济融合程度不断加深，对人工智能治理越来越聚焦到具体的场景上。

1. 人工智能治理实质化进程加速推进

当前，各国人工智能治理侧重各有不同，但整体上呈现加速演进态势，即从初期构建以“软法”为导向的社会规范体系，开始迈向以“硬法”为保障的风险防控体系。

欧盟从伦理向监管稳步推进，欲引领全球人工智能监管规则。2021 年 4 月 21 日，欧盟公布了《人工智能法》草案，是全球范围内首部系统化规制人工智能的法律，细化了人工智能四级风险框架，重点针对高风险系统作出规制，并提出了较为完善的监管配套措施。这是欧盟继发布《可信人工智能伦理指南》（2018）和《人工智能白皮书—通往卓越和信任的欧洲路径》（2020）后的又一重要举措，

标志着全球人工智能治理从伦理原则等软性约束，迈向全面且具有可操作性的法律规制阶段。

美国强调审慎监管以促进创新发展。最早由 2019 年行政令《保持美国在人工智能领域的领导地位》奠定了美国在人工智能治理方面以强化全球领导地位为核心的总基调。美国 2019 年提出《算法问责法案》，要求对“高风险”的自动决策系统进行影响评估；2020 年，美国参议院发布《国家生物识别信息隐私法案》，在个人隐私数据保护的基础上，针对人工智能赋能的生物信息识别进行了隐私保护；2021 年 5 月，美国《算法公正与在线平台透明度法案》从用户、监管部门和公众三个主体维度提出算法透明的义务要求。2021 年 7 月，美国政府问责局发布人工智能问责框架，以确保人工智能系统的公平、可靠、可追溯和可治理等。

中国软法硬法双兼顾，齐头推进人工智能治理。原则伦理层面，国家新一代人工智能治理专业委员会继 2019 年 6 月发布《新一代人工智能治理原则——发展负责任的人工智能》之后，于 2021 年 9 月发布《新一代人工智能伦理规范》，旨在将伦理道德融入人工智能全生命周期，积极引导全社会负责任的开展人工智能研发与应用活动。**法律进程方面，**我国尚未出台人工智能相关的统一法律，但是 2021 年 11 月正式实施的《个人信息保护法》，与《网络安全法》《数据安全法》共同形成了治理人工智能底层要素的坚固法律体系。此外，地方层面积极探索，深圳于 2021 年 7 月出台《深圳经济特区人

工智能产业促进条例（草案）》，助力人工智能产业健康发展。

与此同时，英国、法国、日本、韩国等国家也针对人工智能治理开展了相关工作。英国强调人工智能规范发展并推动 AI 教育及人才培养，在《人工智能：未来决策的机会和影响》（2016）、《英国人工智能发展的计划、意愿和能力》（2018）、《新兴技术宪章》（2021）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。法国通过专家研讨、学术辩论等方式深化对人工智能伦理问题的认识。日本、韩国等从制造业智能化转型和新兴技术应用等发展的角度关注人工智能伦理。

2. 典型场景化治理各有侧重加速落地

人工智能治理的复杂还体现在其应用场景的多样化和差异化。在不同场景下，人工智能技术的应用深度和影响各有不同，典型场景的治理成为各国的工作重点，特别聚焦于自动驾驶、智慧医疗和人脸识别等领域。

自动驾驶领域，德国率先制定伦理准则及框架法案，各国加紧部署分级分类监管。德国于 2017 年推出《自动驾驶伦理准则》，于 2021 年 5 月通过《自动驾驶法》草案；2021 年，英国讨论修改《公路法》，引入自动驾驶汽车在高速公路上安全使用的新条款；我国于 2021 年 5 月由国家互联网信息办公室会同有关部门起草了《汽车数据安全管理若干规定（征求意见稿）》，向社会公开征求意见。

智慧医疗领域，伦理原则逐步得到发展，监管层面注重规制医疗器械准入。美国 FDA 在 2019 年《人工智能医疗器械独立软件修正监管框架（讨论稿）》的基础上，于 2021 年 1 月发布了《基于人工智能/机器学习的医疗器械软件行动计划》，部署人工智能医疗器械软件监管行动。欧盟出台医疗器械条例（MDR），要求自 2021 年 5 月新的医疗器械申请合规性证书。中国于 2021 年 6 月发布了《人工智能医疗器械注册审查指导原则（征求意见稿）》，并推动人工智能医疗器械行业有序发展。

人脸识别领域，全球各国迈入隐私保护和信息数据安全的强监管时代。欧盟在 2021 年 4 月出台的《人工智能法》草案中将人脸识别纳入高风险分类等级，10 月欧洲议会投票通过决议，呼吁全面禁止基于人工智能生物识别技术的大规模监控。中国于 2021 年 11 月实施《个人信息保护法》，与 8 月最高人民法院出台的人脸识别相关的司法解释针对性地规制人脸信息处理。美国州或地方层面通过立法禁止政府机构在公共场所使用人脸识别技术。英国于 2021 年 9 月发布《新兴技术宪章》，指出要合法道德的使用人脸识别等技术。

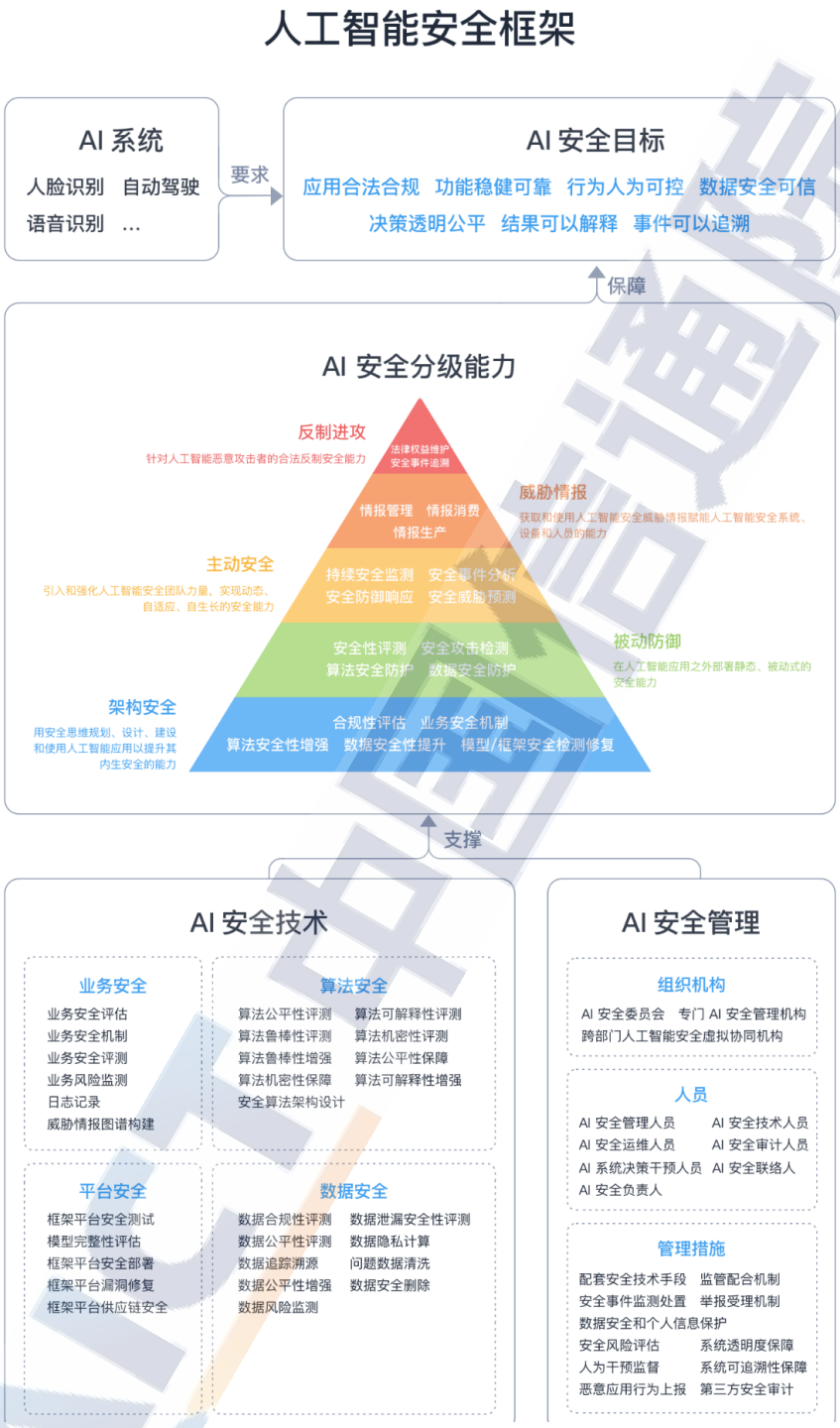
（三）人工智能安全框架成为有效防范风险的关键指引

为有效防范人工智能技术应用带来的安全风险，保障事关国家安全、经济命脉、社会稳定等的人工智能系统的安全，亟需提出人工智能系统安全体系，为行业逐步提升人工智能安全能力提供有效

指引。人工智能安全框架是从人工智能安全保护需求出发，将人工智能安全技术体系和人工智能安全管理体系进行有机融合，构建的人工智能安全整体体系设计与规划，对维护国家人工智能安全和网络安全具有重要意义。

1. 人工智能安全框架逐渐形成雏形

人工智能安全框架需包含安全目标、安全能力、安全技术和安全管理四个维度，如图 4 所示。这四个防护维度基于自顶向下、层层递进的方式指导企业构建人工智能安全防护体系。其中，设定合理安全目标是保障人工智能应用安全的起点和基础，安全能力是实现安全目标的有效保障，安全技术和安全管理是安全能力的支撑和体现。



来源：中国信息通信研究院

图 4 人工智能安全框架[17]

[17] http://www.caict.ac.cn/kxyj/qwfb/ztbg/202012/t20201209_365680.htm

2. 分类分级成为框架构建的新方向

分类分级成为全球人工智能治理的新风向。欧盟《人工智能法》草案、美国《算法问责法案》、加拿大《自动化决策指令》、中国《关于加强互联网信息服务算法综合治理的指导意见》等世界主要国家的法律法规和政策文件均提出建立人工智能系统或算法分类分级管理方式的要求。然而，上述法律法规或仅提出分类分级管理要求或采用列举典型人工智能系统的方式描述分级方式，缺少分级原则，没有可遵循的分级方法和流程，无法适用于快速涌现的新型人工智能应用,亟需提出人工智能分类分级体系，明确分类分级原理以及便于实际操作的分类分级要素和方法。

按照分类管理、分级保护的思路，本白皮书提出了以下人工智能分类分级建议。根据人工智能系统的自主程度的不同，将人工智能系统分为辅助人类智能系统、人机混合智能系统和完全自主智能系统三类。依据人工智能系统的重要性和危害程度，将人工智能系统分为中低风险智能系统、高风险智能系统和超高风险智能系统三级。其中每一类人工智能系统均可进一步分为三级。

（四）可信人工智能已成为落实治理要求的重要方法论

根据《可信人工智能白皮书》[9]，面对人工智能引发的全球信任焦虑，发展可信人工智能已经成为全球共识。**可信人工智能是从产业维度出发，落实人工智能治理要求的一整套方法论，是人工智能**

治理和产业实践之间的桥梁。图5给出了可信人工智能的总体框架。

来源：中国信息通信研究院

图 5 可信人工智能总体框架

1. 可信理念逐渐深入到人工智能全生命周期

可信人工智能从学术界提出，到各界积极研究，再到产业界开始落地实践，其内涵也在逐步的丰富和演进。可信人工智能已经不再仅仅局限于对人工智能技术、产品和服务本身状态的界定，而是逐步扩展至一套体系化的方法论，涉及到如何构造“可信”人工智能的方方面面，包括企业内部管理、研发、运营等环节，以及行业相关工作，将相关抽象要求转化为实践所需的具体能力要求，从而提升社会对人工智能的信任程度。

2. 企业已成为实践可信人工智能的主要力量

企业作为人工智能技术研发和创新应用的一线，需要直面人工智能信任挑战，主动开展自律自治工作，充分发挥企业能动性落实人工智能技术、产品和服务的可信要求。2018 年以来，谷歌、微软、IBM、旷视、腾讯等众多国内外企业纷纷推出了企业人工智能治理准则，并形成相应部门机构推动落实治理责任。另外，企业也在积极探索以实践可信为核心理念的人工智能治理模型，IBM、微软、华为、京东等国内外企业发布多个人工智能可信工具，以帮助人工智能产品在研发过程中提升安全性、鲁棒性、可解释性、公平性等可信能力，并通过开源生态凝聚开发者宣传可信理念。

3. 行业组织推进打造人工智能安全可信生态

可信人工智能的实现不仅仅是企业单方面的实践和努力就能够完成的，更需要多方参与和协同，最终形成一个相互影响、相互支持、相互依赖的良性生态。**在标准制定层面**，2017 年以来 ISO/IEC、IEEE、SAC/TC 28/SC 42 等国内外标准组织已率先布局通用可信人工智能标准。2021 年 4 月，我国在人脸识别场景下的国家标准《信息安全技术人脸识别数据安全要求》面向社会公开征求意见。**在行业自律层面**，中国人工智能产业发展联盟 2019 年发布了《人工智能行业自律公约》，随后在 2020 年发布了《可信 AI 操作指引》，并公布了首批商用人工智能系统可信评估结果，涉及 11 家企业的 16 个人工智能系统，为用户选型提供了重要参考。当前，正在联合业

界共同编制《可信人工智能研发管理指南》等，以期推动人工智能研发源头的安全和可信。

四、总结与展望

我国人工智能技术和产业已经取得了长足的发展，我们相信“十四五”期间，人工智能技术创新将进一步加快，产业规模持续扩大，将涌现出一批发展潜力大的优质企业和产业集群，成为引领经济高质量发展的重要引擎。

追求技术创新、聚焦工程实践、确保可信安全逐渐成为未来人工智能发展的重要方向。回顾近十年的人工智能发展历程，不难发现技术创新与工程实践相辅相成，算法和算力突破后带动了工具体系的发展，工具的成熟进一步又支撑了技术落地应用。当前，人工智能已广泛应用于人们日常生产、生活的方方面面，对其安全可信品质的需要已经提升到前所未有的高度，推动人工智能可靠可控的发展成为全球共识。站在“十四五”的开端，我们期待人工智能技术持续改善，人工智能产业及应用能在下一个五年内蓬勃且健康地发展。

一是在新技术不断探索的同时，更加注重通过工程化的方式释放技术红利，并且确保安全可信。人工智能企业能否快速赋能各行各业，响应多样化需求，其关键因素在于企业的工程化能力。同时，安全可信技术的需求越发重要，当前围绕着数据保护已经催生了大

量从事隐私计算技术的企业，未来围绕着人工智能稳定性、公平性等方面的技术也将会形成重要的力量。

二是在产业智能化进程中，传统行业的参与程度将越来越深入，甚至会主导整个产业的发展进程。产业发展重心已经开始从“人工智能+”向“+人工智能”转变，随着传统行业数字化进程的提升，将提供海量的数据和丰富的应用场景，为人工智能的应用打开新的空间。这些传统行业或领域中，人工智能渗透率更高的机构将会对整个领域内其他机构输出人工智能相关解决方案。

三是人工智能治理工作将越发关键，事关人工智能持续健康发展，统筹治理和发展成为必需。治理工作不仅切实关系到人工智能日常应用问题，也已上升为国际间竞争与合作的重要议题。面临世界各国各地区不同文化背景、不同发展程度，如何有效的开展人工智能治理实践是重要的挑战。我国政府、行业组织、企业等已在人工智能治理方面率先开始探索，将安全可信的理念融入到人工智能的全生命周期中，未来也将涌现出更多的实践范式。

参考文献

[1] Han X, Zhang Z, Ding N, et al. Pre-Trained Models: Past, Present and Future[J]. 2021.

[2] Tamkin A, Brundage M, Clark J, et al. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models[J]. 2021.

[3] Liu H, Wang R, Shan S, et al. What is Tabby? Interpretable Model Decisions by Learning Attribute-based Classification Criteria[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, PP(99):1-1.

[4] 张钹, 朱军, 苏航. 迈向第三代人工智能[J]. 中国科学: 信息科学, 2020, 50(9):22.

[5] 先进计算发展研究报告[R]. 北京: 中国信息通信研究院, 2018.

[6] Yu Sun, Shuohuan Wang, et al. ERNIE 3.0: Large-scale Knowledge enhanced Pre-Training for Language Understanding and generation. [J]arXiv preprint arXiv:2107.02137, 2021.

[7] Yuqi Huo, et al. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training[J] arXiv preprint arXiv:2103.06561, 2021.

[8] Danny Hernandez, Tom B. Brown. Measuring the Algorithmic Efficiency of Neural Networks[J]. arXiv preprint arXiv:2005.04305,

2020.

[9] 可信人工智能白皮书[R]. 北京：中国信息通信研究院, 2021.

[10] Eykholt, K., et al. Robust physical-world attacks on deep learning visual classification[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1625-1634, 2018.

[11] Liu T., et al. Algorithm-dependent generalization bounds for multi-task learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, pages 227-241, 2016.

[12] He F., et al. Control batch size and learning rate to generalize well: Theoretical and empirical evidence[C]. In Advances in Neural Information Processing Systems, pages 1141-1150, 2019.

[13] Tu Z., et al. Theoretical analysis of adversarial learning: A minimax approach[C]. In Advances in Neural Information Processing Systems, pages 12280 – 12290, 2019.

[14] 人工智能安全白皮书[R]. 浙江：浙江大学-蚂蚁集团金融科技研究中心，2020.

[15] 郭锐. 人工智能的伦理和治理[M]北京：法律出版社，2020.

[16] 人工智能治理白皮书[R]. 北京：中国信息通信研究院, 2020.

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：010-62309514

传真：010-62304980

网址：www.caict.ac.cn