cic 工信安全

人工智能算力基础设施安全发展白皮书

国家工业信息安全发展研究中心

2022.11

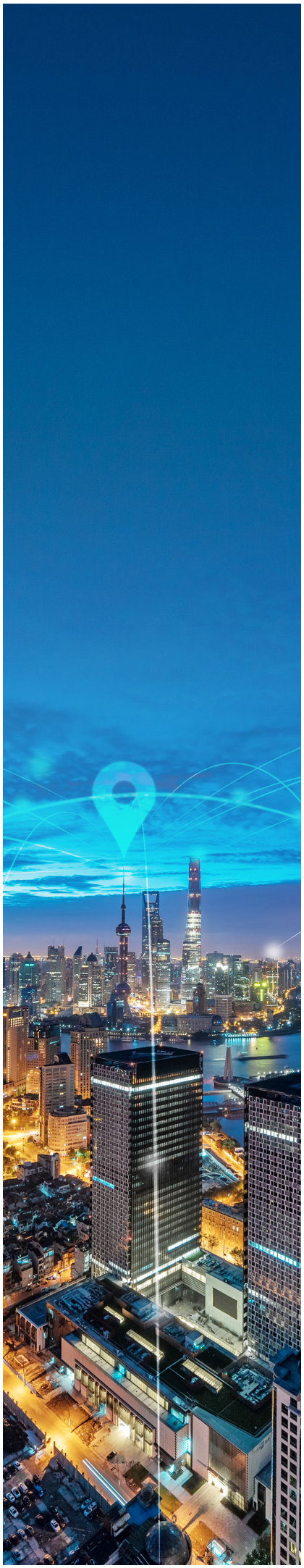
PREFACE
前言

人工智能作为新一轮科技革命和产业变革的重要驱动力量，正以其强大的赋能作用与各领域加速融合，应用范围不断拓展，行业渗透率快速提升。与此同时，人工智能产业对算力的需求已成为影响其发展与应用的核心因素之一。近年来，人工智能算力基础设施建设取得了长足进展，通过构建人工智能算力网络、保障大模型算力、提供普惠算力，在助力人工智能生态建设、保障人工智能产业持续发展方面发挥着越来越重要的作用。

在当前复杂的安全形势下，人工智能算力基础设施由于其属性多样、节点复杂、用户数量多以及人工智能自身脆弱性等特性，在应用过程中已暴露出数据模型窃取、对抗样本攻击、节点不可信等安全问题，带来了更加复杂多样的安全风险，使得人工智能算力基础设施在建设和运营过程中面临着更为严峻的安全挑战，同时影响了用户对人工智能算力基础设施的安全信任，阻碍了其算力资源潜力充分释放。

在此背景下，国家工业信息安全发展研究中心撰写《人工智能算力基础设施安全发展白皮书》，研究人工智能算力基础设施安全的内涵与体系架构，分析人工智能算力基础设施安全管理现状并提出发展建议，旨在为人工智能算力基础设施安全建设提供方法思路，为用户选择和使用安全的人工智能算力基础设施提供判别依据，为人工智能产业健康、持续发展提供决策参考。

由于人工智能算力基础设施正处于快速发展时期，我们对其认识还有待进一步深化，白皮书中难免存在需要改进之处，敬请广大读者指正，也欢迎业界同仁共同参与完善，为人工智能安全发展提供助力！



CONTENTS

目录

第一章
人工智能算力基础设施安全发展背景与意义

随着人工智能融合发展与应用的步伐不断加大，人工智能已经成为世界各国竞争角逐的焦点。人工智能算力基础设施作为人工智能发展的重要基础性资源，在快速发展的同时正面临日益复杂的网络信息环境和安全形势，遇到算力节点不可信、算力信息泄露、对抗样本攻击、数据投毒、模型窃取等各种类型的安全挑战。推动人工智能算力基础设施安全发展，建立安全的人工智能算力基座，对于推动构建人工智能安全生态具有重要的现实意义。

一 人工智能算力基础设施安全发展的背景

1.1 人工智能算力基础设施建设是国家算力战略的重要内容

（1）人工智能算力基础设施建设是“东数西算”战略落地的重要支撑

数字经济时代下，伴随着数字技术向经济社会各领域全面渗透，全社会的算力需求愈加迫切，算力已成为推动经济数字化转型发展的核心生产力。据中联数据预测，社会对算力的需求将长期保持每年20%的快速增长。2022年2月17日，国家发改委、中央网信办、工业和信息化部、国家能源局联合印发通知，同意在京津冀地区、长三角地区、成渝地区、粤港澳大湾区启动建设全国一体化算力网络国家枢纽节点。至此，我国“东数西算”工程全面启动，正式进入实施阶段。作为开启算力经济时代的世纪工程，“东数西算”战略明确把算力中心作为基建投资对象进行布局，旨在整合优化全国算力资源，通过构建一体化新型算力网络体系，将东部算力需求有序引导到西部，充分利用西部地区气候、能源、环境等优势提供算力生产，推动东西部地区算力供需互补与匹配，以满足数字经济带来的强劲算力需求。

随着人工智能技术的不断发展与应用，人工智能已成为数字经济发展的新引擎，人工智能算力也将成为未来算力的主要力量。据华为《智能世界2030》报告数据预测，未来10年，通用算力将增长10倍，而人工智能算力将增长500倍。据中国信通院预计，到2023年，新增算力中人工智能算力将达到70-80%，成为未来算力的主要增长方向。人工智能算力基础设施是针对人工智能算力需求特点而专门设计的算力基础平台，其智能计算效率要远高于传统数据中心。围绕我国经济数字化发展带来的算力需求结构调整，积极推动人工智能算力基础设施建设，更加契合“东数西算”高效配置算力资源以实现集约化、绿色化发展的整体战略部署理念。可以说，加强人工智能算力基础设施建设，将对“东数西算”战略目标的实现产生长远的影响，也将成为建好“东数西算”工程，促进国家数字经济发展的重要基础。

（2）人工智能算力基础设施建设是新型基础设施建设的关键环节

随着我国经济发展加快数字化、网络化、智能化转型，新型基础设施建设对于推动经济高质量发展发挥着至关重要的作用。2020年4月20日，国家发改委首次明确了新型基础设施的范围，认为新型基础设施是“以新发展理念为引领，以技术创新为驱动，以信息网络为基础，面向高质量发展需要，提供数字转型、智能升级、融合创新等服务的基础设施体系”，主要包括信息基础设施、融合基础设施和创新基础设施三个方面。其中，以人工智能、云计算、区块链等为代表的新技术基础设施，以数据中心、智能计算中心为代表的算力基础设施被明确列入信息基础设施的范畴。具体而言，新型基础设施建设对于推动经济发展具有以下重要作用：一是新型基础设施建设可带动传统基础设施的数字化、网络化、智能化改造升级，提高传统基础设施效率；二是新型基础设施建设可有效推动供给侧结构性改革，推动产业结构升级；三是新型基础设施建设有利于充分发挥数字经济潜力，进一步促使交通、医疗、金融、制造业等领域实现生产方式和商业模式革新。我国高度重视新型基础设施建设，中央经济工作会议和政府工作报告多次提出要加快新型基础设施建设，《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》对新型基础设施建设作出了明确部署，强调要“增强数据感知、传输、存储和运算能力”“强化算力统筹智能调度”。总体而言，新型基础设施建设对于夯实经济高质量发展基

础、保障经济行稳致远具有重要而深远的意义。

人工智能算力基础设施是新型基础设施建设的重要领域。一是人工智能新型基础设施建设已经成为新型基础设施建设的核心支撑。首先，人工智能新型基础设施建设可助力其他领域新型基础设施建设转型升级。人工智能作为新一代信息技术的代表，具有很强的通用性和基础性,对支撑5G基站、大数据中心、特高压、城际高速铁路和城市轨道交通、新能源汽车充电桩、工业互联网等新基建的智能化升级具有重要促进作用。其次，人工智能新型基础设施建设与其他领域新型基础设施建设相结合，可形成叠加效应，加速赋能传统产业数字化转型。二是人工智能算力基础设施是人工智能新型基础设施建设的重点。随着我国对人工智能算力的需求急速增长，人工智能算力基础设施作为人工智能发展的重要底层基础设施被推向发展新高地。国家工信安全中心《新一代人工智能算力基础设施发展研究报告》指出，我国人工智能算力基础设施正在加快建设，但仍落后于应用的需求。据IDC调研显示，超过九成的企业正在使用或计划在三年内使用人工智能，其中74.5%的企业期望在未来可以采用具备公用设施意义的人工智能算力新型基础设施。综上所述，随着数字经济的发展，人工智能算力基础设施建设已成为人工智能新型基础设施的重点建设内容，也成为提升新型基础设施建设水平、发挥新型基础设施赋能效应的重要手段。

1.2 人工智能算力基础设施是人工智能产业发展的重要基石

（1）保障大模型算力，塑造“大模型+大算力”基础能力

当前，随着人工智能在多元化场景加速落地，长尾场景应用对人工智能算法的泛化性、通用性提出了较高要求，已成为人工智能产业化发展中面临的重要挑战，而人工智能大模型成为解决这一挑战的重要手段。人工智能大模型是“人工智能预训练大模型”的简称，包含了“预训练”和“大规模”两层含义，即模型在大规模数据集上完成预训练后无需调整，或仅需要少量数据的微调，就能直接支撑各类应用。大模型解决了以往传统的小模型通用性差、部署性弱的问题。目前，全球人工智能企业为抢占技术前沿纷纷布局大模型的研发与应用。国外OpenAI、谷歌、微软、Facebook、NVIDIA等机构纷纷构建大模型平台，并形成了GPT-1、GPT-2、GPT-3、BERT、Switch Transformer等典型的大参数量预训练模型。我国华为、百度、智源研究院等企业和组织也加大人工智能模型研制和应用。2021年4月，华为云发布盘古系列超大规模预训练模型，包括30亿参数的全球最大视觉预训练模型以及千亿参数、40TB训练数据的全球最大中文语言预训练模型。2021年6月，北京智源人工智能研究院发布“悟道2.0”人工智能模型，参数规模达到1.75万亿，是OpenAI公司GPT-3模型参数的10倍，打破了之前由国外预训练模型创造的1.6万亿参数记录。2021年12月，百度正式发布了知识增强千亿大模型“鹏城-百度·文心”，参数规模达2600亿。可以说，全球人工智能已全面进入大模型时代。

大模型训练需要的庞大算力支撑，企业和科研机构自建人工智能算力基础设施较难满足其峰值算力需求。同时，即使企业可以通过自建人工智能算力基础设施提供足够算力，由于计算业务天然存在的波动性，也会导致算力闲置、能源浪费的情况。因此，企业自建人工智能算力基础设施既不符合企业经济性运营要求，也不符合国家“双碳”目标要求。推动人工智能算力基础设施建设，可有效平衡人工智能算力基础设施运行中的算力需求，削峰填谷，充分利用能耗资源，集约化、高效率提供人工智能算力。只有加快推动人工智能算力基础设施发展，构建经济适用、高效环保的人工智能算力基座，塑造“大模型+大算力”的人工智能产业基础能力，才能充分发挥大模型的技术驱动效应，进而推动人工智能产业加速落地。

（2）构建人工智能算力网络，支撑全国人工智能协同发展

随着我国人工智能产业加速集聚发展，区域人工智能算力需求日益庞大而集中，孤立单一的人工智能算力基础设施已经无法满足地方发展对算力的需求。发展人工智能算力网络，可以将各地的人工智能计算中心联接成网，动态实时感知算力资源状态，实现统筹分配和调度计算任务，构成区域内计算资源的统一管理、协同调度及弹性分配网络，在此基础上汇聚和共享算力、数据、应用资源，为区域提供充裕的算力。因此，发展人工智能算力网络是人工智能产业集群化发展的必然结果。只有通过建设人工智能算力网络，促进算力资源的进一步汇聚和协同调度，才能释放出更大的算力能力，在助力人工智能自身技术发展的同时，更好地发挥人工智能赋能效应。

当前，我国在人工智能要素资源禀赋方面存在着东西部产出与需求间的不平衡。东部地区人员密集，数据信息资源丰富，人工智能计算需求大，但存在能源、土地及能耗指标紧张，建设人工智能算力基础设施成本高等问题；西部地区能源丰富、气候适宜，具备人工智能算力基础设施的良好环境，但存在网络带宽小、跨地区数据传输费用高等瓶颈。统筹推进人工智能算力基础设施建设，将全国各地的人工智能算力基础设施进行网络互连、资源共享，形成一体化人工智能算力网络，可将西部的最优算力资源服务于东部的计算需求，在全国层面促进算力和数据等人工智能基础要素的合理流动，推动人工智能产业整体规模扩张和应用水平提升，降低功耗和计算成本，形成算力、数据、能耗平衡发展的全国人工智能协调发展格局。随着我国人工智能赋能效应不断提升，构建全国一体化人工智能算力网络，更好地汇聚和调配全国算力、数据资源，不仅是满足人工智能大规模算力需求、提高西部绿色能源利用率的必然要求，也是我国经济社会高质量发展的必然趋势。

（3）提供普惠算力，降低人工智能创新创业门槛

人工智能创新创业主体主要是面向各个分散应用场景的人工智能企业，其特点是企业数量众多、算力需求大、成本敏感性高。尤其是随着人工智能模型参数量越来越多，人工智能算力成本不断攀升，大大提高了人工智能创新创业门槛。据广发证券研究报告显示，2019年，谷歌推出的BERT大模型拥有3.4亿个参数，训练到目标精度的花费为1.5万美元；2020年，Open AI推出的GPT-3大模型拥有1750亿参数，训练成本达到了1200万美元；2021年，微软和英伟达使用了4480个GPU训练出的拥有5300亿参数的MT-NLG大模型，其训练成本更是高达8500万美元。目前，我国主要的算力基础设施仍以超算中心为主，人工智能计算中心仍处于短缺状态。超算中心立足于科学研究，主要应用于重大工程或科学计算领域的通用科学计算，提供的算力特点是性能高、价格贵，并不十分符合人工智能企业的算力需求。人工智能算力基础设施立足于为大规模人工智能算法和模型研究提供算力支撑。扩大人工智能算力供给，能够降低人工智能算力使用成本，从根本上解决人工智能研发企业、科研机构存在的算力获取困难的问题，降低人工智能创新创业门槛，帮助企业将资源专注于人工智能技术研发和场景应用，支撑更高精度的模型开发及更高质量的应用孵化。

面对人工智能企业算力需求倍增与训练成本高企的挑战，推动人工智能算力基础设施建设，在为大量人工智能企业模型研发提供充裕算力的同时，以规模效应降低公共人工智能算力基础设施的建设和运营成本，为企业提供普惠算力，更广泛的支持人工智能研发和应用，培育产业生态。具体来讲，建设人工智能算力基础设施对于企业创新创业具有以下作用：一是通过汇聚算力资源，发挥算力规模效应，降低人工智能算力服务成本。二是通过将人工智能算力资源以普惠价格开放给人工智能企业、科研机构和高校，让算力服务像水和电一样成为社会公共设施，提高企业使用公共算力服务便利性。

人工智能算力基础设施安全发展的意义

2.1 人工智能算力基础设施安全发展符合国家加强信息技术安全的政策指引

（1）发展安全的人工智能算力基础设施是落实《关键信息基础设施安全保护条例》的现实要求

2021年7月30日，国务院颁布《关键信息基础设施安全保护条例》（以下简称《条例》），不断强化关键信息基础设施的科技伦理治理和科技自立自强要求。《条例》针对关键信息基础设施安全，建立专门保护制度，明确各方责任，提出保障促进措施，开启了我国关键信息基础设施保护的新格局，也为关键信息基础设施安全发展指明了方向。《条例》明确认定：“关键信息基础设施是指公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务、国防科技工业等重要行业和领域的，以及其他一旦遭到破坏、丧失功能或者数据泄露，可能严重危害国家安全、国计民生、公共利益的重要网络设施、信息系统等”。《条例》明确了关键信息基础设施安全保护工作的对象和关键流程，明晰了国家对关键信息基础设施的重点保护、保障和促进措施，划分了关键信息基础设施运营者的责任义务，对于落实和细化网络安全法关于关键信息基础设施运行安全相关规定，有效保障关键信息基础设施安全具有重要的指导意义。

随着人工智能的快速发展和在国民经济发展中的应用不断增强，人工智能算力基础设施已成为引领数字经济、智能产业发展的关键信息基础设施。人工智能算力基础设施可能遇到数据泄露、后门攻击等安全挑战，风险源头众多，风险防控难度较大。而且随着人工智能加速“基建化”，人工智能算力基础设施将加快转变为像水、电一样的基础设施，加速渗入到国民经济的各个方面，一旦出现安全问题，将对国家经济社会运行安全造成威胁。推动人工智能算力基础设施安全发展，对筑牢人工智能算力基础设施安全防线，切实落实《条例》要求，构建国家信息安全屏障具有重要意义。

（2）发展安全的人工智能算力基础设施是支撑人工智能安全发展的有力保障

近年来，人工智能在获得深度应用的同时，也面临愈来愈严峻的人工智能安全风险挑战。一是人工智能自身发展面临潜在严重威胁。针对算力、算法和数据，人工智能在自身发展中面临着算力信息泄露、算力网络攻击、对抗样本攻击、算法后门侵入、模型逆向攻击、数据投毒、隐私泄露等众多潜在安全风险，一旦发生安全事故，势必引起严重的后果，对人工智能技术研发和大规模应用造成严重威胁。二是人工智能无序应用将带来人工智能伦理和隐私安全挑战，加剧社会风险。如智能杀熟算法、人脸识别滥采滥用、性别歧视算法、广告新闻虚假合成、个人隐私挖掘等人工智能技术滥用，将带来隐私伦理方面的安全威胁，严重扰乱市场和社会秩序。

推动人工智能算力基础设施安全发展是提高人工智能安全水平的重要手段。一是可以有效防范与规避人工智能安全风险。人工智能算力基础设施作为人工智能算法运行的基础环境，大部分人工智能安全风险与人工智能基础设施安全紧密相关。提高人工智能基础设施安全发展水平，可以从物理、网络、数据、算法等多领域防范安全风险威胁，推动人工智能整体安全水平提高。二是可以有效赋能人工智能安全检测与评估。人工智能算力基础设施通过提供安全检测和评估工具，可以对算法的公平性、可解释性、鲁棒性、隐私性等进行检测，对人工智能算法安全管理能力进行评价。同时，通过提供隐私计算、攻防博弈等技术工具，人工智能算力基础设施可以帮助提升人工智能隐私保护能力，进

而提高人工智能安全治理水平。

（3）发展安全的人工智能算力基础设施是实现我国人工智能算力自主安全的重要举措

目前，我国在人工智能应用落地方面走在世界前列，但在人工智能算力基础设施的软、硬件技术水平和自主创新程度方面，与国外相比还有一定的差距，人工智能芯片、开源框架、模型、系统工具等关键技术仍高度依赖国外供给。据中国经济信息社江苏中心发布的《新一代人工智能发展年度报告(2020-2021)》显示，在人工智能算力支持方面，IBM、HPE、戴尔等国际巨头稳居全球服务器市场前三位，浪潮、联想、新华三等国内企业市场份额有限；国内人工智能芯片厂商需要大量依靠高通、英伟达、AMD、赛灵思、美满电子、EMC、安华高、联发科等国际巨头供货，中科寒武纪等国内企业发展刚刚起步。面对日益严峻的外部环境，我国人工智能算力基础设施自主创新的重要性日益凸显。如果不能实现人工智能算力基础设施自主创新，将导致我国的人工智能算力发展需要借助国外技术才能开展，也就无法保障我国人工智能算力安全，将导致我国人工智能算力发展面临随时可能被别国“卡脖子”的风险。

人工智能算力基础设施涉及的软硬件技术和产品众多，包括通用芯片、人工智能芯片、传感器、服务器、数据库、操作系统、计算框架等。只有充分保证人工智能算力基础设施软硬件的安全性，打通人工智能算力技术、产品、系统壁垒，才能保障我国人工智能算力基础设施可靠运行，为人工智能提供稳定算力。总体来说，保障人工智能算力基础设施安全，加大对人工智能芯片、人工智能计算框架等软硬件技术领域的布局力度，有利于提升我国人工智能算力自主创新水平，保障人工智能算力可靠供给，从而打造人工智能算力安全基础，更好地发挥人工智能算力在推动经济发展、助力社会治理和拉动科技创新中的重要作用。

2.2 人工智能算力基础设施安全发展是人工智能安全发展的内生需求

在“东数西算”“新基建”等国家战略工程牵引下，人工智能算力产业链条不断延伸，导致在算力生产及数据传输中，人工智能算力基础设施本身面临网络安全、应用安全等多重风险。因此，人工智能算力基础设施在自身发展中具有稳定、可靠、可信、合规的内生安全需求。构建安全的人工智能算力基础软硬件平台,实现主动性和被动性安全防护机制，从指令集、芯片，到服务器、软件系统，打造自主安全算力基座，开创安全可信算力新生态，可以有效促进人工智能算力基础设施应用发展，进而更有力地支撑人工智能产业发展。加强人工智能算力安全体系建设，既是人工智能算力基础设施在面对种类日益繁多的安全风险时的内在发展需求，也成为构筑安全的人工智能算力关键基础设施的必然要求。

鉴于目前人工智能算力基础设施在安全性方面存在的不足，企业在使用算力服务时普遍存在顾虑。企业在使用人工智能算力基础设施时有诸多考量：一是人工智能算力基础设施必须提供安全稳定的算力输出。例如，在智慧城市应用中，必须在7×24小时不间断传输和计算摄像头视频，并防止数据窃取和隐私泄露，以赋能智慧交通和平安城市。二是人工智能算力基础设施必须能够实现对安全威胁的全面感知和溯源，进行实时安全检测和有效对抗处理。因此，推动人工智能算力基础设施安全发展，通过构建弹性灵活、安全可信、交易便利的算力生态，可以为企业提供方便即用、稳定可靠的人工智能算力，打消企业在算力安全方面的顾虑，助力构建人工智能算力安全生态。

第二章
人工智能算力基础设施安全的内涵与体系架构

人工智能算力基础设施是以软硬件基础设施为底层支撑，以算力、数据、算法等资源平台为核心要素，实现算力生产调度、数据开放共享、算法开发调用等功能，支撑人工智能与各领域渗透融合的基础设施体系，从技术维度包含人工智能基础软硬件、算力平台、数据集、算法仓库等部分。人工智能算力基础设施由于涉及层次多、分布范围广、接入设备繁杂、用户数量多等特性，其安全问题也面临多重维度，安全风险来源较为复杂。作为人工智能系统运行的基础载体，人工智能算力基础设施应明确其面临的安全种类和风险来源，建立全面有效的安全防御体系，为人工智能系统安全保驾护航。

一 总体框架

人工智能算力基础设施安全指为人工智能算力基础设施建立和采用的技术和管理层面的安全保护，目的是保护人工智能算力基础设施硬件、软件、人工智能数据模型等不受到破坏、更改和泄露，保障其为人工智能系统提供安全的算力和运行环境。人工智能算力基础设施安全具有三重属性。一是基建属性。作为“基础设施”，人工智能算力基础设施应对其稳定性、可用性、可靠性等自身安全提供保障。二是技术属性。作为“AI算力”，人工智能算力基础设施应对部署在其之上的人工智能系统的运行安全提供保障。三是公共属性。作为“公共设施”，人工智能算力基础设施应对人工智能产品、系统和企业提升安全管理能力、降低安全风险、助力合法合规提供安全服务。

推动AI算力基础设施安全发展应从强化自身安全、保障运行安全、助力安全合规三个方面发力，通过强化自身的可靠性、可用性与稳定性，保障算法运行时的可信度与准确度，提升用户的安全管控力、认可度与合规性等八个领域筑牢人工智能安全防线，打造可信、可用、好用的人工智能算力底座，营造安全、健康、合规发展的人工智能产业生态。

图1 人工智能算力基础设施安全体系架构

二 强化自身安全

强化自身安全是指人工智能算力基础设施应保障自身安全、稳定运行，主要注重以下方面：一是要筑牢传统安全，围绕物理安全、网络通信安全、计算环境安全和数据应用安全等方面提供全方位保障。二是要提升算力网络安全，人工智能算力基础设施作为算力网络的重要节点，应通过加强安全保障提升算力网络的可用性。三是要注重供应链安全，人工智能算力基础设施建设应考虑加强技术自主创新，保障基础软硬件供应链稳定安全。

2.1 筑牢传统安全，保障可靠性

传统安全是人工智能算力基础设施正常运行的基础，人工智能算力基础设施应围绕物理安全、网络通信安全、计算环境安全和数据应用安全等方面提供全方位保障。

（1）物理安全

物理安全指为确保人工智能算力基础设施物理层面稳定运行，对人工智能算力基础设施采取的安全措施。物理安全是人工智能算力基础设施安全的根本保障，直接影响到人工智能系统的可靠性、保密性、完整性、可用性等，主要风险来源包括环境因素和人为因素。其中，环境因素是指由物理环境原因导致人工智能算力基础设施出现故障进而影响人工智能系统正常运行的风险，包括自然灾害、电磁环境影响、物理环境影响、设备故障等。人为因素是指由人为操作影响到人工智能系统正常运行，包括物理攻击、无作为或操作失误、管理不到位等。人工智能算力基础设施物理安全的防护不到位可能导致系统停机甚至损毁等严重后果。例如，2021年3月，欧洲最大的云服务和网络托管服务运营商OVH位于法国的一个大数据中心发生火灾，导致约360万个网站出现故障，超过一万名客户的资料可能受到影响，其中包括法国政府的部分数据。2021年6月，云服务器公司Fastly因为停电导致服务器关机，包括部分亚马逊网站，纽约时报，推特，CNN等等数十个网站停止服务一小时。

人工智能算力基础设施应注重物理安全保障，围绕设备物理安全、环境物理安全、系统物理安全三方面加大安全防护力度。设备物理安全方面，人工智能算力基础设施应具备一定抗强电流、抗电磁干扰、电源适应等能力。环境物理安全方面，人工智能算力基础设施所在环境应具备一定的防火、防雷、防水、防盗、防毁、防静电、温湿度控制、应急供配电等能力。系统物理安全方面，人工智能算力基础设施系统应具备设备信息、软件信息等资源

管理能力，具备设备备份、运行状态监测、告警监测等能力。

（2）网络通信安全

网络通信安全是指人工智能算力基础设施为保障其网络通信系统的软硬件及系统数据不受破坏、维持网络通信系统稳定可靠运行采取的措施。网络通信功能是人工智能算力基础设施的基础功能，直接关系到用户的远程访问使用和数据传输，网络通信中断将直接影响到人工智能算力基础设施对外服务。例如，2021年五月，美国最大的成品油管道运营商Colonial Pipeline因受到网络攻击被迫关闭其美国东部沿海各州供油的关键燃油网络。

人工智能算力基础设施应注重网络通信安全保障，围绕网络结构、访问控制、攻击防范三方面重点进行保障。网络结构方面，人工智能算力基础设施应保障主要网络设备和通信线路冗余，网络性能满足业务高峰需求，各子网络间实现有效隔离。访问控制方面，人工智能算力基础设施应在网络隔离点部署访问控制设备，以最小安全访问原则设置访问控制权限，具备用户访问控制能力。攻击防范方面，人工智能算力基础设施应具备网络入侵检测能力，并在发生攻击行为时提供预警。

（3）计算环境安全

人工智能算力基础设施内部运行环境称为计算环境，由各种设备节点的内部环境共同构成。计算环境安全指为保障人工智能算力基础设施计算环境不被入侵或植入恶意程序采取的措施。计算环境安全是人工智能系统不被攻击的重要保障，不可信的计算环境可能导致人工智能系统被篡改、窃取或破坏，直接影响到人工智能系统正常运行。

人工智能算力基础设施应注重计算环境安全保障，在用户身份鉴别、恶意程序防范、环境安全审计方面重点开展防护。用户身份鉴别方面，人工智能算力基础设施应对登录的用户进行身份识别和鉴别，确保用户身份不被冒用。恶意程序防范方面，人工智能算力基础设施应具备恶意代码检测和预警能力。环境安全审计方面，人工智能算力基础设施应确保设备安全审计，审计覆盖范围全面、记录项目内容完整。

（4）数据应用安全

数据应用安全指人工智能算力基础设施为保护数据在应用过程中不被破坏、更改和泄露而采取的措施。数据应用安全关系到人工智能算法准确性、人工智能系统完整性以及用户隐私保护等多方面。一是数据作为人工智能系统的核心元素，其准确性直接影响到人工智能系统的功能和性能。二是数据作为人工智能公司的重要资产，具有极大商业价值。三是部分数据直接涉及用户个人隐私，若发生数据泄露将直接为用户带来风险。例如，2019年10月，谷歌云服务器上发现了一个汇总了12亿用户个人信息的数据库，该数据库无保护地储存于服务器上，涉及的信息包括社交媒体帐户、电子邮件地址和电话号码等，导致12亿用户个人信息面临泄露风险。

人工智能算力基础设施应注重数据应用安全保障，围绕数据完整性、数据保密性、备份和恢复进行增强。数据完整性方面，人工智能算力基础设施应能够检测数据在存储、传输等过程中是否受到破坏，并能够采取必要的恢复措施。数据保密性方面，人工智能算力基础设施应能够对存储和传输的数据进行加密或采取其他保护措施。备份和恢复方面，人工智能算力基础设施应提供本地和异地数据备份与恢复功能，并且确保备份介质安全。

2.2 提升算力网络安全，增强可用性

人工智能算力基础设施作为算力网络的重要节点，应通过加强安全保障提升算力网络的可用性。人工智能算力网络将各地分布的人工智能算力基础设施节点联接起来，构成多个算力节点间感知、分配、调度人工智能算力的网络，弹性满足全网范围内的算力需求，汇聚和共享数据、模型等人工智能资源，有助于推动构建区域范围乃至全国范围的人工智能产业生态网络。据人民日报报道，截至今年

6月底，我国数据中心机架总规模超过590万标准机架，服务器规模约2000万台，算力总规模超过150EFlops（每秒1.5万京次浮点运算次数），位居全球第二。作为我国算力网络的重要组成部分，人工智能算力基础设施在规模不断扩大的同时，也面临多端运算网络带来的节点不可信、暴露面增多、审计溯源复杂等安全风险。

人工智能算力基础设施建设应从节点可信认证、算力网络管理规范、行为审计溯源等方面加强算力网络安全防护，保障人工智能系统全程可溯、多方安全。节点可信认证方面，由于算力网络接入设备繁杂，为保证接入的每个节点安全可信，应加强节点可信认证，保证每个接入节点在硬件和软件层面实现全流程安全认证，并确保算力网络业务安全接入，实现人工智能算力基础设施算力网络节点全程安全可信。算力网络管理规范方面，为解决多端运算众多用户和设备节点带来的安全风险，人工智能算力基础设施应建立统一安全管理规范，将不同节点纳入统一管理体系，保障算力网络管理安全合规。行为审计溯源方面，为解决算力共享带来的行为审计困难问题，人工智能算力基础设施应建设算力网络协同行为安全记录机制，实现多方算力行为可审计可溯源。

2.3 注重供应链安全，提升稳定性

人工智能算力基础设施建设应考虑加强技术自主创新，保障基础软硬件供应链稳定安全。供应链安全是指人工智能算力基础设施应保障其中央处理器、图形处理器等基础硬件和操作系统、基础框架等基础软件供应稳定和安全。一是安全稳定的供应链直接关系到人工智能算力基础设施能否稳定建设和运营，若技术无法自主研发且供应依赖于少数供应商或供应国，一旦基础软硬件断供将对人工智能算力基础设施造成重大打击。二是基础软硬件的安全性直接关系到人工智能算力基础设施安全性，安全的供应链能够防止基础设施软硬件被供应方植入后门或存在其他未知风险。例如，2022年8月美国出台《芯片和科学法案》，特别要求接受其财政补助的芯片厂商不得在中国新建、扩展先进制程工艺的半导体厂。该法案的生效和实施，将影响全球芯片产业链供应链的优化配置和安全稳定。

人工智能算力基础设施建设应考虑建立自主标准规范体系，加强技术自主创新，打通技术壁垒，采用具有自主知识产权的通用处理器、人工智能专用处理器、高性能内存、传感器等基础硬件和操作系统、数据库、人工智能框架等基础软件，保障供应链安全，提升基础设施运行稳定性。

三 保障运行安全

保障运行安全是指人工智能算力基础设施应提供安全的运行环境，保障人工智能系统的机密性和完整性。作为第三方服务提供方，人工智能算力基础设施以云服务、按需索取的形式为用户提供了简便易用、专业化的模型训练算力服务；为了保障算力服务的安全可信，下列安全问题和建设要点应当予以进一步重视。

3.1 保护数据模型不被窃取，保障机密性

现如今的人工智能行业中，高性能算法模型的研制是核心支柱；同时，与模型训练相配套的高质量数据是极为珍贵的数字资产，获取难度非常高，还常常包括人脸、指纹等受到严格监管的敏感隐私数据。因此，保障人工智能模型与数据在算力基础设施运行环境中的安全，已经成为整个人工智能生态链条的关键环节；如果无法保障数据模型机密性，将严重打击人工智能模型厂商使用人工智能公共算力基础设施的积极性，不利于人工智能生态的进一步蓬勃发展与价值实现。

安全技术层面，应着力研制人工智能算力基础设施内置用户模型保护技术，重点防御窃取攻击。神经网络模型推理的输出结果往往会包含着模型自身的许多特征信息，可被用于窃取模型及其训练数据的相关信息，从而逆向构建出具有相近功能的模型，进而展开成功率更高的白盒攻击。为了在训练环节中实现对隐私数据和模型的保护，目前通常采取差分隐私计算、对模型参数或者输出结果进行近似处理、利用联邦学习等等方法。在这些相互独立的基础防护技术之上，应进一步着眼于整体人工智能运算环境构建安全防护方案。例如，在大规模人工智能算力基础设施中，通过高性能加密技术、容器完整性保护、身份与权限分级严格管理等手段，凭借高性能算力支撑构建全程可信赖的安全运行环境，有效保护数据和模型所有者对其核心资产的所有权。

安全制度层面，应当完善人工智能算力基础设施内部安全管理规章体系。人工智能算力基础设施作为提供算力云服务的底座，承载着价值极高的算法模型和海量的训练数据；安全管理制度不完善、人员队伍建设编制不明确、算力设施从业人员安全意识薄弱、人员身份职责与权限划分不清晰等等人员管理层面可能存在的相关问题，都将为人工智能算力基础设施运行过程中的模型和数据机密性造成威胁。因此，应当围绕技术和管理规范并重的核心思想，加强人工智能算力基础设施的内部信息安全规范体系构建。例如，可参考大数据中心等相近功能机构信息安全制度规范，建立明确的责任分工机制和授权机制，配备符合条件的人员，加强定期培训，严格确保相关人员按照既定政策、程序和权限履行职责，保障数据、模型在使用、销毁等各环节不被窃取。

3.2 防范数据模型遭受恶意攻击，保障完整性

出于各种不法或恶意目的，人工智能算法模型在运行过程中往往会遭受多种形式的恶意攻击，导致模型产出错误的运行结果，进而带来应用风险。因此，除开发者在模型内部自行设计安全增强模块之外，人工智能算力基础设施所提供的算力服务环境也应针对主流恶意攻击风险，提供相应的预警和响应机制。

（1）数据投毒攻击检测与防御

数据投毒攻击是出现最早的，针对深度学习模型的攻击方式之一，Biggio等研究者于2012年提出这一概念。由于基于深度学习的人工智能模型在大量训练样本数据基础上训练而来，攻击者可在训练数据中注入精心设计的恶意攻击样本，污染数据，

进而干扰模型结果的准确率。2017年 11 月底至2018 年初，谷歌Gmail邮箱遭遇了至少 4 次大规模恶意攻击，垃圾邮件制造者试图通过将大量垃圾邮件提交为非垃圾邮件，试图让垃圾邮件评估分类器失效。阿里巴巴公司在2019年也受到了一种模型偏斜的数据投毒攻击：攻击者前后发送了两轮攻击流量，第一批低级攻击流量的目的是为了干扰入侵检测系统，发送集中于模型边界的大量恶意样本，使模型黑白样本分布不均匀，进而导致模型偏斜；而第二轮攻击流量是一批对抗攻击流量，原本就在模型边界的对抗样本，在模型偏斜之后，能够更容易地绕过入侵检测系统的过滤。受到攻击的模型，性能会遭到全面削弱，便于攻击者从事下一步不法行为，因此数据投毒攻击也被称为可用性攻击。

针对数据投毒攻击，目前主要可从模型开发者和深度学习框架两个角度展开应对。从模型开发者角度，应避免小范围来源的训练样本在总样本规模中占据过大的比重；使用AB测试、暗启动、回溯测试等方式，在发布模型的更新版本之前将其与最新的稳定版本进行比较；创建可信赖的基准数据集。从深度学习框架角度，可通过框架提供的鲁棒性增强算法，对模型进行强化开发；同时，可通过Fuzz Testing安全测试模块检测不同类型的模型输出结果、错误行为，进而分析模型的安全性。

（2）后门攻击检测与防御

在传统安全领域，后门指能够绕过软件的安全性控制，通过隐蔽的通道获取对程序或系统的非法访问权的攻击。2017年，Gu等人发表的BadNets和Liu等人发表的TrojanNN 提出了在基于深度学习的人工智能模型中也存在后门。模型后门是攻击者在训练中向深度神经网络模型植入的隐藏模式，目的是诱导模型错误决策。后门攻击可贯穿深度学习系统整个生命周期，不会干扰正常样本输入，此时模型将给出正常的推理结果和准确率；但攻击者可以利用特定的输入，或在输入上附加特殊的模式（称为trigger）激活模型中的后门，使模型产生符合攻击者预期的输出结果，从而造成危害。例如，在计算机视觉领域内的自动驾驶任务中，摄像头捕捉的图像是输入数据中非常重要的一环；攻击者在神经网络中植入后门以后，通过在摄像头捕获的图像上叠加Trigger，特定情境下触发后将导致系统对行驶路线做出错误的预判决策，在实际环境中可能导致车毁人亡的惨剧。

由于涉及深度学习系统的不可解释性问题，后门的生成机制和trigger激活后门的机制并不透明。这也导致，目前的应对措施通常是针对特定的攻击

手段进行特化防御，尚不存在一种通用的解决方案。后门攻击的防御可以从数据和模型两方面着手进行：对于数据，可以通过输入转换、输入过滤等方式进行筛查；对于模型，则可以通过模型净化,模型检测方法加以核查。其中，前者要求有海量良性数据集，而后者需要强大的计算资源。因此，未来有望通过大规模人工智能算力基础设施提供的强大运算能力，从模型安全核查角度提升后门攻击的应对能力。

（3）对抗样本攻击检测与防御

对抗样本攻击最早于2013年提出，是指通过特定的算法向人工智能模型输入干扰数据构成对抗样本，利用人工智能模型的信息反馈机制发起攻击，导致模型算法在正常运转中输出一个错误的结果，进而影响人工智能模型决策的置信度。具体而言，对抗样本攻击可分为两类：一是数字对抗攻击，在数字空间生成对抗样本攻击模型，例如能够对图片实现bit级的极细微控制要求；二是物理对抗攻击，能够在真实物理场景下的实现对AI模型的攻击。近年来，对抗样本攻击技术呈现出如下特点：从白盒到黑盒，门槛越来越低；从数字到物理，威胁越来越严重。对抗攻击技术一旦被用于人脸识别、自动驾驶以及智能监控等领域的，将严重威胁AI时代智能应用在关键行业的落地与生态信任的建立。例如，2021年1月，瑞莱智慧公司在测试中，利用人脸对抗样本技术生成了眼部的干扰图像，并将其打印出来贴在眼镜的框架上，攻击者戴上该眼镜就能破解19款常见安卓手机和十余款金融和政务类App的人脸识别锁。

针对对抗样本攻击事件，目前模型开发者方常用的防御手段包括两种方式：一是模型重训练，即将已知的对抗样本纳入到训练数据集和测试数据集中重新训练模型；二是防御性蒸馏，即将原模型的知识迁移至参数较少、结构简单的替代模型中，降低模型对输入扰动的敏感度。目前，随着深度学习框架的不断发展，可以凭借框架内置的安全测试模块进行验证，使用不同的对抗样本攻击方法模拟恶意攻击者对输入数据添加各种扰动，以评测AI模型在实际的对抗样本攻击下的鲁棒性,并通过模型增强工具弥补模型缺陷、增强鲁棒性。同时，针对开发者层面完善反对抗样本攻击的困难，可以在人工智能算力基础设施层面的AI系统应用环境中引入部署攻击检测引擎，通过深度学习实时检测并提取攻击特征，及时发出警告反馈。

四 助力安全合规

助力安全合规是指人工智能算力基础设施应对人工智能产品、系统和企业提升安全管理能力、降低安全风险、助力合法合规提供安全服务。一是在安全检测方面，围绕数据集完整性、准确性及算法公平性、鲁棒性、可解释性等重点领域，提供安全检测工具。二是在安全评估方面，通过提供自评估工具、引入第三方评估等手段帮助用户对其人工智能产品的安全问题及合规风险开展评估及认证。三是在安全增强方面，通过提供可信审计、隐私计算等工具，帮助用户增强安全合规性。

4.1 提供安全检测能力，助力用户加强安全管控力

人工智能算力基础设施应围绕数据集完整性、准确性以及算法公平性、鲁棒性、可解释性等重点领域，为用户提供安全检测工具，帮助用户提升安全风险识别和管理能力，在数据准备、模型训练、系统运行等全流程检查人工智能产品的安全风险。例如，在算法安全检测工具方面，华为提出了MindArmour安全可信工具包，针对模型鲁棒性、用户隐私风险、数据漂移等功能提供了相应检测工具。鲁棒性检测工具方面，提供了多种对抗样本生成、检测和防御方法以及攻防评测指标，可从恶意攻击角度测评模型以及非恶意扰动角度评测模型鲁棒性。隐私泄露评估工具方面，提供了模型逆向攻击方法和成员推理方法，从攻击角度测评模型泄露隐私的风险。数据漂移检测工具方面，提供了时序数据和图像数据的概念漂移检测，检测在线模型的输入数据是否发生漂移现象并提前预警。在算法可解释性工具方面，华为提出了基于昇思MindSpore可解释AI工具箱MindSpore XAI，为用户提供对黑盒模型决策的解释和评价标准，提高用户对模型的理解和信任。MindSpore XAI已为表格类数据、图像类数据提供了十余种解释方法，并提供了一套对解释方法效果评分的度量方法，从多种维度评估解释方法的效果，帮助用户选择最适用于特定模型和场景的解释方法。MindSpore XAI也包含了一系列可解释模型，帮助用户打造原生可解释产品。目前

MindSpore XAI已在金融领域上线金融智能营销可解释业务和风控业务，为相应业务提供语义级解释结果，帮助业务满足金融行业合规要求。

4.2 提供安全评估能力，助力用户提升安全认可度

人工智能算力基础设施应通过提供自评估工具、引入第三方评估等手段帮助用户对其人工智能产品的安全问题及合规风险开展评估及认证，增强该产品的安全认可度。评估方式可以是为用户提供问卷或清单式的检查，对用户的人工智能算法安全管理能力进行评价，并通过完善管理手段帮助其提升算法安全性。评估工具可在人工智能系统开发、部署的早期阶段就帮助企业评估其安全管理能力并帮助用户建立完善且具有针对性的管理制度，通过持续地执行和监督，促进制度的落实，确保负责任地开发、部署和维护人工智能系统。例如，2022年9月，国家工业信息安全发展研究中心发布《人工智能安全风险管理体系研究》报告，在分析梳理人工智能面临的安全风险和国内外主要管理措施的基础上，提出我国人工智能安全风险管理体系构建的原则与思路，建议从规则引导、技术攻关、评估认证三大维度构建可落地的人工智能安全风险管理体系，提出从制度、组织、数据、算法、性能、安全六大维度建设人工智能安全风险管理框架。下一步，将在此基础上，依据人工智能融合发展应用领域的特色和实际情况，进一步深化框架，编制安全风险管理指南和评估规范，帮助人工智能算法提供方、应用方和监管方提升安全风险管理能力。

4.3 提供安全增强能力，助力用户增强安全合规性

安全增强是指人工智能算力基础设施通过提供一定服务，帮助用户增强人工智能系统的安全合规性。安全增强服务可由人工智能算力基础设施自身提供，也可由第三方安全服务商提供，相关工具集成至人工智能算力基础设施，在人工智能系统开发、运行等阶段，用户可以选择不同方向、不同程度的安全增强服务对自身人工智能系统进行安全增强，进一步提高人工智能系统安全合规性。安全增强工具主要包括可信审计工具、隐私计算工具等。

（1）通过提供可信审计工具保障监管合规

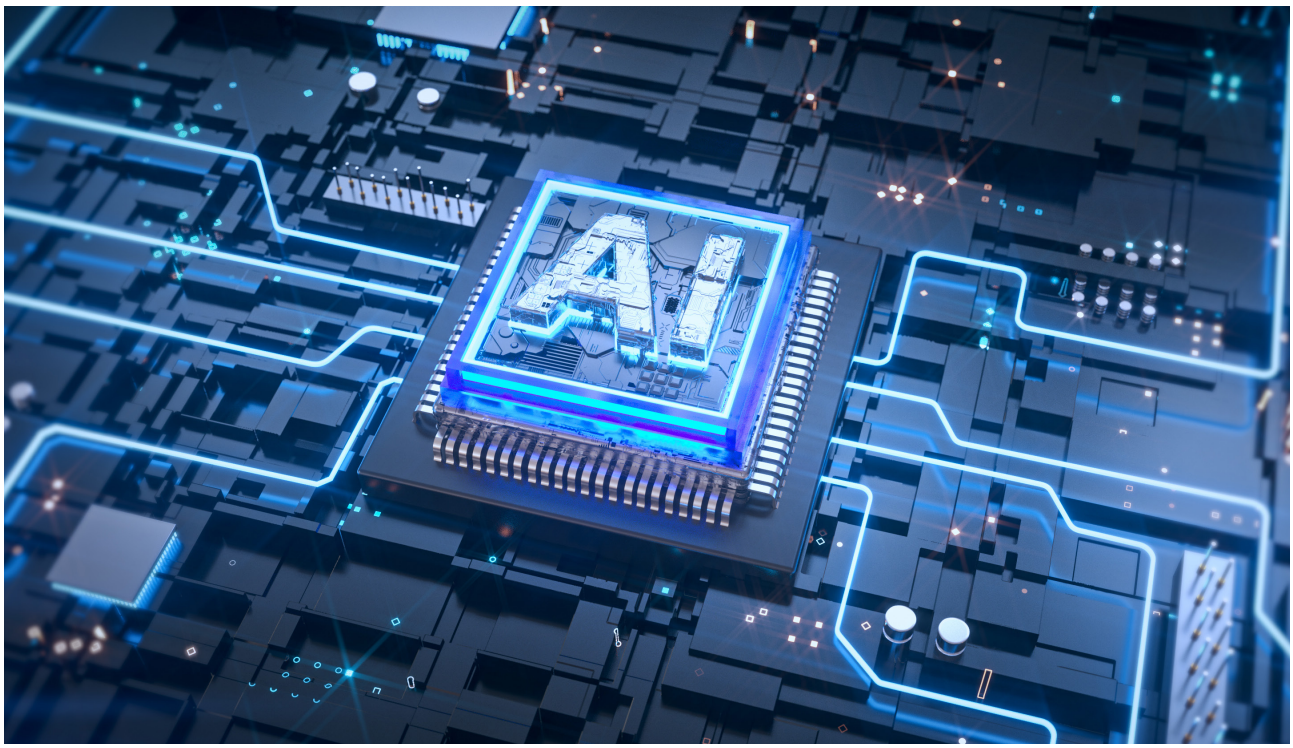
人工智能系统日志在监管合规与用户隐私方面具有关键作用，人工智能算力基础设施应为人工智能系统全生命周期中的日志审计提供安全保障。一是可信审计有助于监管溯源。人工智能系统日志是保证人工智能系统全生命周期可追溯性的重要手段，记录了对算力环境的运维操作和对用户数据的访问等敏感行为，是对人工智能系统实施监管和行为溯源的重要方式。二是可信审计能够增强用户信任。对于用户来说，人工智能算力基础设施可信的日志记录有助于其及时掌握管理人员的行为操作，有效缓解用户使用人工智能算力基础设施的隐私顾虑。

为保障人工智能系统日志自身面临的安全风险，防止攻击者或管理人员未经授权访问隐私数据后删除、篡改相关日志记录，人工智能算力基础设施应为人工智能系统全生命周期的日志审计提供安全可信保障，为监管机构提供安全可信的责任溯源，缓解用户对利用公共算力环境进行人工智能训练和推理的数据模型隐私顾虑。例如，华为基于密码学中的Merkel树构建了不可篡改的日志记录系统，为用户提供了可信审计工具。在系统运行时，人工智能模型厂商在记录、保存人工智能模型的运行日志过程中，基于Merkel树技术同步生成人工智能系统日志的完整性证据，构建可验证的审计路径来保障日志完整性。通过将完整性证据保存至安全存储区域，构建了日志可信审计基础，实现了对攻击者、运维人员篡改日志文件行为的检测和预警。

（2）通过提供隐私计算工具增强数据算法安全

人工智能算力基础设施应为用户提供数据、算法等关键信息不泄露的保障手段。隐私计算是指在使用数据、算法的全流程计算操作中，通过隐私计算理论将隐私信息进行处理，在保护用户数据不

对外泄露的前提下实现数据分析计算，主要包括联邦学习、多方安全计算、可信执行环境、多方中介计算等技术。人工智能算力基础设施采用隐私计算技术可实现数据和模型分离，充分保护用户隐私安全。例如，在数据模型保护工具方面，2020年8月，鹏城实验与奇安信联合建立AI靶场，基于协同计算实现了跨多个计算集群的分布式协同隐私计算作业。该AI靶场基于“数据不动程序动、数据可用不可见、分享价值不分享数据、保留所有权释放使用权”的隐私保护理念，构造了一个可信的执行环境，研究人员可在环境中安全使用数据，但无法带走数据。若用户不愿上传自身数据到AI靶场，也可通过鹏城众智协同计算平台使用本地语料数据与AI靶场数据进行联合训练或微调。在联邦学习方面，华为提出了MindSpore Federated联邦学习框架，解决了保护用户隐私以及模型训练过程中的数据孤岛问题。该框架采用精度无损的安全聚合算法、轻量级的局部差分隐私加噪算法、高精度维度选择算法SignDS等方法，实现了用户原始数据不出本地的前提下，支持多方联合建模，实现了隐私计算。MindSpore Federated是业界首个面向大规模用户、企业级场景的联邦学习框架，可支持千万级无状态设备与数据中心联合部署，赋能全场景智能应用，目前已在半监督用户画像挖掘、无监督用户群组构建等业务中大规模部署上线。在模型加密方面，华为MindSpore安全团队研发了模型混淆技术以进一步增强人工智能模型安全性，可防止高权限管理员或攻击者从内存中窃取模型。相对于传统加解密、密态计算、基于可信执行环境的方案等模型保护方法难以平衡模型推理性能和机密性保护，模型混淆技术通过模型结构混淆、算子混淆和权重混淆等技术对明文形式的AI模型的结构和权重进行加扰保护，使得混淆之后的模型不会泄露原模型的结构和权重信息，同时对模型的推理性能影响较小，具有保护程度灵活，性能开销小，部署范围广（不依赖特定操作系统或可信硬件环境）等特点。该技术目前已被Harmony OS、MLKit等产品成功应用，在人脸、检测等模型的测试结果比较中，该方案的推理性能为传统加解密方案的数十倍。



第三章
人工智能算力基础设施安全管理现状

当前，人工智能向着多领域运用、多行业触及、多资源牵涉的方向不断深入发展，无论是用户需求的外在广度还是模型训练复杂性的内在深度，都对算力提出了更具挑战性的要求。在内生与外生双重需求的共同作用下，近年人工智能算力基础设施建设迎来高潮。热度之下，人工智能算力基础设施安全稳定成为人工智能实现高质量可持续融合发展的重要保证。由于当前人工智能算力基础设施发展仍处于初期阶段，尚未形成体系化的安全监管框架，专门的安全标准、技术体系、评估规范、监测和检查手段仍在建设发展过程中，但可通过针对人工智能安全、算力基础设施安全的相关管理手段分析当前推动人工智能算力基础设施安全发展的主要做法与实践进展。目前，全球多个国家和地区已在人工智能算力基础设施领域迈出了探索步伐，在政策、标准、技术三个层面上取得了一批实践探索成果，推动全球人工智能算力基础设施与新兴领域安全战略政策举措融合向实发展。

一 政策引导方面
各国不断细化明确相关安全管理规定以提供安全发展指引

全球各主要国家围绕人工智能算力基础设施的“基础设施”与“人工智能”两大重点属性范畴，逐步厘清人工智能算力基础设施安全要素组成部分，在治理政策上逐步深化顶层设计，完善和细化法规体系、组织机构、治理原则，实现治理效能的有机提升，引领持续安全发展。

1.1 关键信息基础设施安全要素逐步明确

以人工智能算力基础设施为代表的新型关键信息基础设施，不仅应用范围广，还具有涉及功能模块多、领域交叉程度高的特点。因此，需要重点围绕其所涉及的各项功能与组成模块，梳理明确安全要素，进而有针对性地改进管理组织架构和监管法规体系。

（1）美国明确关键信息基础设施管理机构和监管体系

美国是全球最早针对关键基础设施展开法规保护的国家。迄今为止，美国共计发布了20余项关键信息基础设施相关的保护政策，重点关注关键基础设施的分类管理，以及如何应对层出不穷的新型网络攻击。

在关键基础设施的分类管理方面，美国启动时间早、持续关注程度高。1996年，克林顿签署13010号行政令《关键基础设施保护》，组建了“关键基础设施保护委员会”，规定了8个关键基础设施行业，正式开启了美国对关键基础设施的法治监管探索。2013年美国发布第21号总统令《关键基础设施安全和恢复力》，取代2003年发布的7号总统令，最终确定16个行业并得以固化。

同年，美国发布第13636号行政令《提升关键基础设施网络安全》，标志着关键基础设施保护的网络安全时代到来，针对关键基础设施的安全保护研究重心逐渐向网络攻击事件倾斜。2021年，美国总统拜登签署行政命令及备忘录，旨在实现关键基础设施网络防御措施现代化。今年年初以来，在俄乌冲突进一步加剧国际局势不稳定的背景下，《2022年关键基础设施网络事件报告法》增加了关键基础设施实体在遭遇网络事件和因勒索软件攻击而支付赎金时的两项强制性报告义务，进一步增强了政府对关键基础设施网络安全风险事件规模和特征的感知能力。

此外，美国还重视通过成立具体管理与支撑机构的方式，跨领域统筹资源投入，以更灵活的组织机构形式保障政策实施，例如建立了国家人工智能

研究资源（NAIRR）工作组。2022年5月，该工作组发布中期报告，提出了如何构建、设计、运作和管理美国信息基础设施的愿景；该报告中关于计算资源和系统安全、用户访问控制两方面的阐述，集中反映了目前美国对于人工智能算力类关键基础设施建设需求和安全保障的前沿关切。

（2）欧盟由组织和制度角度发力提高风险管控能力

针对关键信息基础设施领域的保护工作，欧盟的主要思路是引入风险分级管理思想，推行基于风险管理策略的安全治理，向以消减脆弱性为主的新安全理念转变，对重点应用进行重点保护。

2004至2006年间，欧盟启动“欧盟关键基础设施保护规划”，审议通过《欧洲关键基础设施保护计划》绿皮书，开始重点关注关键基建安全保护与发展。2011至2017年间，陆续出台《网络与信息安全指令》《关键信息基础设施领域的物联网安全基线指南》等多项政策，推动成员国间的安全战略协作和信息共享。

近年来，欧盟对新型网络安全事件的关注度日益上升，例如2020年发布的《欧盟网络安全战略》，将增强关键信息基础设施的保护水平和恢复能力作为未来五年网络安全领域的核心工作。目前，欧盟已推出了“欧洲关键基础设施保护计划”，在这一框架性架构下，构建欧洲关键基础设施保护参考网络，各国实验设施和实验室通过该网络开展信息共享和交流协作，提升关键基础设施安全事件响应能力。

总体而言，欧盟侧重由发挥成员国组织协作制度优势的角度出发，从四个方面着力展开保护。一是引导政府公共部门和私营企业部门、成员国之间展开安全战略协作和信息、经验的沟通共享。二是通过发展欧盟内部的信息共享与预警机制，加强安全事件的感知、检测与响应能力。三是通过制定有效的应急预案、应急响应演练等多种方式，强化关键信息基础设施的安全事件防御能力，不断巩固防灾减灾性能。四是推动各成员国关键安全标准的构建与欧盟体制框架下的兼容共通发展。

（3）中国关键信息基础设施安全保护法规体系日趋完善

2016年，我国《网络安全法》正式通过，关键信息基础设施安全保护制度被首次提出。第三章中专门设置“关键信息基础设施的运行安全”专节，对关键信息基础设施的运行安全进行明确规定，指出国家对公共通信和信息服务、能源、交通、水利、金融、公共服务、电子政务等重要行业和领域的关键信息基础设施实行重点保护；在此之上，进一步对于关键信息基础设施安全保护的基本要求、分工以及主体责任等问题作出法律层面的总体安排。《网络安全法》明确规定，基于国家强制标准GB 17859-1999《计算机信息系统安全保护等级划分准则》发布以来一系列国家标准中提出的网络安全等级保护制度基础上，对关键信息基础设施实行分等级重点保护。由此，关键信息基础设施的安全标准上升为法规要求，为我国关键信息基础设施系统安全发挥了更为显著的保驾护航作用。

2021年9月，我国正式施行《关键信息基础设施安全保护条例》，自上承接《网络安全法》要求，进一步明确关键信息基础设施的认定与安全保护范围，对关键信息基础设施保护系列制度要素作了具体规定，涵盖了关键信息基础设施运营者责任义务、保障和促进、法律责任等诸多方面。《条例》体现出如下的鲜明特点：第一，精准划分责任层级体系，中央层级统筹网信、公安、重要行业和领域主管监管部门，自上而下实现统筹、监管、主管、地方、运营者和服务机构一体化联动；第二，从制度、技术、预警评估等方面内防安全漏洞，与外防违法犯罪攻击、加强安全保卫相结合；第三，突出重点领域，将能源和电信行业摆在更为突出的特殊位置优先保障，并强调要为其他行业领域提供重点保障；第四，因地制宜，发挥国家力量牵引攻关重点自主化工程，提升关键信息基础设施安全可信可控整体水平。

1.2 人工智能安全风险治理得到高度关注

人工智能应用的高速发展，在近年来人工智能算力基础设施建设的热潮中得到了直观体现。然而，只有在安全风险治理原则与框架下保障算法产业有序良善发展、确保算法不作恶，算力基础设施才能更好地发挥出高效的正面价值，更好地成为助推产业生态发展的力量倍增器。

（1）美国推进治理原则和风险管理框架出台，高度重视算法问责

针对联邦机构和其他组织如何负责任地使用人工智能的问题，美国提出了一套问责框架。2021年6月，美国政府问责局（GAO）发布了《人工智能：联邦政府和其他机构的责任框架》。GAO认为，人工智能的运行和操作对用户是不可见的，透明度缺乏将为监管和审查带来阻力。同时，人工智能有可能放大与公民自由、道德和社会差异相关的偏见和担忧。因此，GAO强调应重视人工智能问责，以确保相关技术和系统的公平、可靠、可追溯和可治理，并引导美国政府和所有参与设计、开发、部署和持续监测人工智能系统的机构与企业负责任地使用该技术。框架分为治理、数据、性能和监测4大原则，并在每个原则下提出了关键做法、关键问题和问责程序等内容。其中治理原则是指通过建立管理、运营、监督实施的流程，帮助个实体促进问责制和负责任地使用人工智能系统；性能原则帮助实体确保人工智能系统产生与预期一致的结果目标。数据原则是指确保数据来源和处理过程中的质量、可信性、代表性；监督原则是确保随着时间推移，人工智能系统仍具有可靠性和相关性。该框架为今后创立相关立法、出台政策设定了原则和方向，是美国在人工智能治理方面取得的重要进展。

同时，针对人工智能的全生命周期安全风险，美国发布了风险管理框架。2022年3月，美国国家标准与技术研究院（NIST）发布《人工智能安全风险管理框架（初稿）》，旨在帮助指导人工智能风险管理框架的开发，解决人工智能产品、服务和系统在设计、开发、使用和评估过程中的风险，并本着自愿原则进行使用，计划在2023年初发布人工智能风险管理框架1.0版本。框架将人工智能风险管理活动分为映射、测量、管理和治理四项，目的是指导政府、行业等相关部门单位开展人工智能设计、开发、部署和使用。其中，映射指通过识别上下文环境，确定与人工智能相关的风险；测量指对识别的风险进行分析、定量或定性评估，并跟踪其影响；管理活动根据预测的影响对风险进行排序并采取行动，以最大化利益和最小化不利影响；治理活动指在内部培养和实施风险管理文化，以确保风险和潜在影响得到有效且一致的识别、测量和管理。

（2）欧盟率先开展人工智能领域立法尝试，重点关注高风险应用

在人工智能监管与治理的探索中，欧盟逐步形成了以《可信人工智能伦理指南》为核心的可信人工智能政策规范体系。2019年4月，欧盟人工智能高级别专家组发布《可信人工智能伦理指南》，在一些领域开展了实践探索和意见反馈工作，并在指南中列出了可信人工智能评估清单。《可信人工智能伦理指南》从顶层设计的角度提出了构建可信人工智能框架的3个要素（合法、符合伦理、稳健）、4项原则（尊重人的自治、预防伤害、确保公平、具有可解释性）和7项关键要素求（实现人类自主和监管、确保技术稳健和安全、重视隐私和数据治理、注重透明性、注重多样性非歧视、实现环境友好和社会福祉、建立问责制）。

随后，2021年4月，欧盟委员会发布了全球首份人工智能立法提案《欧洲议会和理事会关于制定人工智能统一规则（人工智能法）和修订某些欧盟立法的条例》（以下简称“欧盟人工智能立法提案”），提出了对人工智能系统实行分级监管的思想，以全面、系统地提出人工智能治理法则。该提案从“维护欧盟的技术领先地位，并确保欧洲人民可以从按照欧盟价值观、基本权利和原则人工智能中受益”的原则出发，采取基于风险的监管思路，在区分“禁止类人工智能”和“高风险类人工智能”的基础上，提出了具体目标：确保投放到欧盟

市场并使用的人工智能系统是安全的，并尊重现行法律中的基本权利和欧盟价值；确保法律上的确定性，以促进对人工智能的投资和创新；加强对适用于人工智能系统的基本权利和安全要求的现有法律有效执行和治理；促进针对合法、安全和可信赖的人工智能应用的开发，促进欧盟单一数字市场的发展。该提案重点对高风险人工智能系统监管进行了较为详细的规定，明确了利益相关方的各项义务。第一，在人工智能系统开发过程中，建立风险管理系统、进行数据质量检验、设计用户告知和日志追溯功能等。第二，在人工智能系统系统首次运营前或系统升级迭代后，及时开展合规性评估，并在欧盟委员会建立和管理的大数据库中进行备案登记。第三，在人工智能系统投入使用过程中，建立与人工智能风险级别相匹配适应的售后监测系统与人为监督机制，对人工智能应用中的风险进行监控预警，在发现可能的风险后召回系统处理并通知相关监管机构。第四，在人工智能系统发生故障或严重事故时，应立即采取补救措施，并及时向相关监管部门报告。

（3）中国确立人工智能治理原则，打造算法安全综合治理格局

在人工智能产业和应用高速铺开、广泛渗透各领域的背景下，我国不断致力于通过建设治理原则等多种方式，构建人工智能安全综合治理新格局。2019年2月，国家新一代人工智能发展规划推进办公室宣布成立国家新一代人工智能治理专业委员会，全面开展人工智能治理方面的政策体系、法律法规和伦理规范研究。为应对人工智能发展过程中引发的安全、隐私、公平等问题，我国已提出人工智能治理原则。2019年6月，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》，提出为发展负责任的人工智能，各相关方应遵循和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八项原则。 2021年9月，新一代人工智能治理专业委员会发布《新一代人工智能伦理规范》，强调将伦理道德融入人工智能全生命周期，促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题，提出人工智能特定活动应遵守的伦理规范包括管理规范、研发规范、供应规范、使用规范4大类，共18项具体伦理要求，为从事人工智能相关活动的自然人、法人和其他相关机构等提供了伦理指引。

目前，我国已初步形成了涵盖顶层设计、管理规定、落地实施的算法治理体系。在顶层设计方面，2021年9月，国家互联网信息办公室等九部门发布了关于加强互联网信息服务算法综合治理的指导意见》，提出了“利用三年左右时间，逐步建立治理机制健全、监管体系完善、算法生态规范的算法安全综合治理格局”的目标。在管理规定方面，2021年8月，国家互联网信息办公室等四部门联合发布《互联网信息服务算法推荐管理规定（征求意稿）》，2021年11月正式审议通过，自2022年3月起实行。该规定主要规范“生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类”算法服务，包含总则、信息服务规范、用户权益保护、监督管理、法律责任五个部分，在算法备案、定期审核评估、算法优化、用户沟通等方面都提出了具体和落地的管理措施。在落地实施方面，2022年3月，互联网信息服务算法备案系统正式上线运行，备案主体须在提供服务的十个工作日内通过该系统进行备案。2022年8月，国家互联网信息办公室公开发布了境内互联网信息服务算法的备案信息，在备案系统中对算法的名称、基本原理、运行机制、应用场景、目的意图等进行了公示。

此外，我国业已针对互联网信息服务中运用到的深度合成、智能推荐等具体算法，在规定层面展开了严格管理的探索。例如在深度合成领域，2022年1月，国家互联网信息办公室发布《互联网信息服务深度合成管理规定（征求意见稿）》，规范深度合成相关服务活动，明确了深度合成服务商的责任和义务，并提出了“制定并公开管理规定和平台公约”“定期审核、评估、验证算法机理”“加强训练数据管理”“对深度合成内容进行显著标识”等多项具体做法。

三 标准建构方面，围绕算力基础设施安全与人工智能安全的标准制定工作稳步推进

针对关键基础设施运行的通用性、基础性安全标准，各国行业主管和标准化建设机构主要以方法、建议、指南等形式出台技术指导方案，进而结合实践运行经验得失，逐步构建和完善安全标准体系。目前，人工智能系统的安全标准建构已经得到广泛重视，在世界各国取得了一批初成体系的建设成果；同时，针对新兴的人工智能集约化算力基础设施领域，各主要国家和国际组织主要参照云服务和传统关键信息基础设施的安全标准进行指导，业已纷纷展开规划部署，面向高质量发展迈出了探索细分领域标准的步伐。

2.1 针对网络攻击等共性风险来源，不断强化关键基础设施通用化保护标准

尽管以人工智能算力基础设施为代表的各类关键基础设施具体涉及的应用领域和场景各不相同，但却均面临着一批相同的安全风险问题，其中最为显著的是在网络化高速推进背景下层出不穷的各类新型网络攻击。针对共性风险的防范，美国、欧盟和欧洲国家，以及中国都展开了积极的探索。

（1）美国重点关注网络安全事件，建立兼具防御和威慑能力的安全框架与标准体系

为了在基础安全标准的基础上更加有效地应对针对关键基础设施发起的网络攻击事件，美国在鼓励通过国内国际化标准机构发挥社会化力量推出具体安全标准条款、完善安全标准体系的同时，也重视通过政府引领的方式，构建标准化的安全框架与保护技术体系。

在面向计算与数据中心类的关键基础设施的具体标准条款方面，美国国家标准协会提出了ANSI/TIA942-a2014《数据中心电信基础设施标准》。该标准聚焦数据中心的可靠性，主要关注的方面为计算与数据类基础设施的通信线缆布置和网络建构规范，提出了一系列基础设施冗余规定，同时也关注电力供应、散热冷却、网络传输弹性等问题。ANSI/BICSI002-2014《数据中心设计和实施最佳实践标准》则重点考量了数据中心在规划、设计、建

造和调试等工作中的主要方面，以及防火，IT和维护等关键安全要素。

在安全保护技术体系方面，美国于2009年启动“全面国家网络空间安全行动计划”，核心目标是监测针对政府网络的入侵行为，保护美国联邦政府的网络空间基础设施安全，“国家网络空间安全保护系统”（NCPS），是该计划中一项覆盖全美关键信息基础设施的标准化安全技术体系。NCPS分阶段不断拓展防护技术和防护范围，旨在为美国互联网侧安全态势感知构建四项关键能力：入侵检测、入侵防御与对抗、安全事件分析、信息高效共享。

在标准化安全框架方面，2014年，美国发布了《改善关键基础设施网络安全的框架》。该框架定义了一套应用于关键基础设施安全的风险管控流程。为支撑该框架的落地执行，美国能源部和国土安全部针对关键信息基础设施开发了网络安全能力成熟度模型，将能力成熟度具体划分为四个级别，用于指导运营者对其信息系统、工控系统等信息资产进行安全评估，并通过划分十个安全域的方式，分别从内部网络安全实践的实现情况为代表的方法论层面和安全实践的制度化程度为代表的管理层面对组织安全能力进行评估。2017年12月，美国国家标准技术研究院发布了《提供关键信息基础设施网络安全路线图（草案）》。在这一路线图的规划中，该机构提出了十二个重要领域，例如：网络供应链可靠性与风险、权限与身份评估管理、数据与信息的隐私安全、有效应对网络攻击事件等，在未来的工作中将持续重点关注、开展跨部门跨领域协同推进。

（2）欧盟和欧洲国家围绕关键信息基础设施的识别和保护展开多元探索，推行基于风险管理策略的安全治理

欧盟和欧洲国家高度重视关键信息基础设施的识别工作，在梳理识别关键信息基础设施的基础上，通过分类管理的方式，着眼于具体应用领域和实际安全短板问题，结合风险管理策略展开安全治理。

2014年，欧洲网络与信息安全局发布了《识别关键信息基础设施服务和资产的方法论》，在该文件中提出了识别关键信息基础设施中服务和资产的一系列方法，致力于通过分类管理的方式对基础设施涉及方加强安全保障。此后，欧盟主管机构就关键信息基础设施领域开展公私合作、安全事件演练、风险评估、信息共享和建立控制措施等方面的标准化发展，推出了《保护关键信息基础设施的考量、分析和建议》《数字服务提供商实施最低安全控制措施技术指南》等技术指导文件。在互联网基础设施、工控系统等具体细分领域，欧盟也陆续发布了相关安全规定。例如，2017年11月，欧洲网络与信息安全局发布了《关键信息基础设施领域的物联网安全基线指南》，围绕物联网垂直应用领域，指导欧洲在关键信息技术设施领域如何应用物联网；该报告分析了物联网的安全需求、威胁态势、风险趋势，构建了安全基线，围绕六类常见的应用安全短板问题提出切实可行的对策。

值得一提的是，英国通过2016年生效的《国家网络安全战略》，制定了一套不同于欧盟和美国做法的关键信息基础设施识别定义标准：既不是某一类具体的企业或机构，也不是某一种具体的基础设施，而几乎是完全抽象的、概念化的界定基础设施，从数字经济影响力、数据资源特性等维度，将英国关键基础设施划分为五类：已取得极大成功且在研发或知识产权具备很强优势的重要企业、个人信息数据拥有者、媒体等高威胁目标、顶级数字经济提供商、保险投资和专业咨询组织等。

具体而言，针对数据中心类基础设施，欧盟目前已提出了EN50600系列标准，涵盖了数据中心基础设施运转的所有方面，包括实体固件设计、运维管理、关键绩效指标、具体技术报告等组成部分，不仅关注了电力、冷却和网络通信等硬件的构建方案，还为运营和管理、安全、能耗和可持续性管理提供了建议。标准委员会正根据此标准开发数据中心成熟度模型，以使数据中心运营商能够根据最佳

实践评估其能源和可持续性管理。此外，该标准还被 国际标准组织ISO选为ISO数据中心标准的基础，目前该标准已直接作为ISO技术规范 ISO/IEC TS 22237 的第 1 至 7 部分加以发布。

（3）中国积极统筹建设关键基础设施安全标准，分阶段构建分级安全保护体系

围绕着不断完善的关键基础设施领域安全保护法律法规体系，我国积极建设标准统筹组织并发挥其标准制定的协调引领作用，着眼于不同时期的发展水平和风险类型特点，分阶段逐步推进网络安全等级保护制度的建设。

2002年，我国成立全国信息安全标准化技术委员会（“信安标委”）作为网络安全国家标准的统一技术归口组织，负责统筹协调组织信息安全领域的标准制定活动。我国于2007年发布了《信息安全等级保护管理办法》，随后在标准机构的筹划牵头下，陆续发布了GB/T 22239-2008 《信息安全技术 信息系统安全保护基本要求》、GB/T 22240-2008《信息安全技术 信息系统安全保护定级指南》等一系列具体实施标准，标志着我国在关键基础设施安全标准领域迈入了分级保护1.0标准体系时代。

随着时代的发展，关键基础设施的概念不断向云计算、人工智能、大规模数据资源、移动互联等方面拓展，进入分级保护标准体系2.0时代。分级保护标准体系不断完善，如更新GB/T 22239-2019《信息安全技术 信息系统安全保护基本要求》、GB/T 22240-2020《信息安全技术 信息系统安全保护定级指南》，以及面向数据中心专门发布与修订的GB 50174-2017《数据中心设计规范》。

目前，信安标委积极贯彻落实《网络安全法》《关键信息基础设施安全保护条例》等法律法规要求，围绕安全保障体系建设各维度，重点关注边界识别、保护要求、控制措施、保障指标、应急体系、检查评估以及供应链安全、数据安全、信息共享、监测预警等重要方面。例如，针对与日俱增的云计算服务需求与数据安全重要性，我国目前启动了对《信息安全技术 云计算服务安全能力要求》的修订工作，安全能力要求增加覆盖涉及数据保护的相关各功能，如线上管理运维平台、网络访问途径、功能接口调用等，保护客户业务数据在网络传

表1 ISO和IEEE组织制定的人工智能安全标准

项目	描述
ISO/IEC TR 24028	可信人工智能概述
ISO/IEC TR 24027	AI系统和AI辅助决策中的偏见
ISO/IEC 23894	人工智能风险管理
ISO/IEC 23053:2022	使用机器学习的人工智能系统框架
ISO/IEC TR 24029-1:2021	人工智能-神经网络强健性评估总览
ISO/IEC DTR 5469	人工智能-功能安全与AI系统
IEEE P2840	负责任的AI许可标准
IEEE P2841	深度学习评估的框架和过程
IEEE P2863	人工智能组织治理的推荐实践
IEEE IC20-027	人工智能和生命科学的负责任创新
IEEE P2802	基于人工智能的医疗设备的性能和安全评估标准

资料来源：国家工业信息安全发展研究中心整理

输过程中的完整性、连续性、安全性。未来，我国将继续加快关键信息基础设施标准研制与试点推广工作。

2.2 国际机构及各国围绕实际人工智能发展现状，加快人工智能系统及具体应用的安全标准建设

目前，人工智能产业分布与具体应用情况在各国家和地区之间存在着较为明显的差异，因此在国际标准化机构着力围绕技术组成要点推出人工智能安全发展共性标准的同时，各国也正围绕自身实际产业应用特点和治理思路展开具体的人工智能安全标准建构探索。

（1）国际标准化组织围绕行业和领域前沿关切，积极布局专门工作组推动标准建设工作

各大国际标准化组织面向人工智能及相关领域的标准创制，着眼于人工智能发展中的安全风险管控问题，成立了一系列专门委员会及子委员会，着力推进标准体系的建构。

ISO和IEC组织面向人工智能及相关领域的标准创制，着眼于人工智能发展中的安全风险管控问题，成立了一系列专门委员会及子委员会，如ISO/IEC JTC 1/SC 42人工智能委员会、ISO/IEC JTC 1 SC 27信息和网络安全及隐私保护委员会等等。电气与电子工程师协会IEEE围绕人工智能系统和应用全生命周期中的伦理问题，于2016年便成立了专门的工作组，设立P7000系列标准项目，重点关注和解决人工智能技术和伦理考量交叉点的具体问题。总体而言，已基本涵盖了人工智能系统总体和流程考量以及面向部分具体应用领域的细则。

（2）美国重视发挥社会化力量作用，鼓励引导探索非限制性的人工智能安全发展原则

美国在人工智能安全标准领域延续了一贯的行业标准建设思路，即重视发挥市场化力量，引导企业、学术机构和标准化团体组织建设人工智能标准；在联邦官方层面，并不针对人工智能产业和技术的发展做过多强制性标准指导要求，而是着重通过成立相应工作组的方式为相关研究部门提供资源、人才等领域的协调、保障与支持，并在标准创制完成之后与国家政策进行融合。就目前而言，得益于于技术和人才汇集的优势，美国在国际标准化组织ISO、国际电工委员会IEC、电气与电子工程师协会IEEE等负责人工智能标准制定的国际标准委员会中，处于实质性领导角色。

针对人工智能国家标准的重要关切，2016年美国国家人工智能研发战略计划确定了多个关键领域的标准化需求：软件工程、度量标准、安全性、可用性等等。美国国家标准与技术研究所（NIST）于2022年3月发布了标准性框架文件《人工智能安全风险管理框架（初稿）》，旨在帮助指导人工智能风险管理框架的开发，解决人工智能产品、服务和系统在设计、开发、使用和评估过程中的风险，本着自愿原则供各方使用。此外，美国国家标准协会等团体机构也积极参与标准制定工作，发布了诸如ANSI/CTA-2089.1-2020《医疗保健领域人工智能应用的定义和特征》等标准条款。

（3）欧洲强调以人为本的发展战略，围绕监管法律建立高风险人工智能技术安全应用标准

欧盟希望通过制定更加严格的规范最大限度减小人工智能风险，特别注重在消费者保护、防止不公平商业竞争、保护个人数据等方面加强立法，完善与法律法规相配套的具体技术标准体系。

2018年4月，欧盟发布纲领性文件《欧洲人工智能战略》，明确以人为本的人工智能发展方向，确保符合人类价值观追求始终是人工智能应用发展建设的首要考虑因素，推动人工智能朝着符合欧盟价值观的方向发展。2021年4月，欧盟发布《人工智能协调计划2021年修订版》，建立人工智能监管与协调法律框架，推动可信人工智能发展。同月，欧盟正式颁布《关于人工智能的统一规则(人工智能法》，要求各开发和实现高风险人工智能系统的公司应当遵循国际和欧洲标准化机构所组织发布的相关领域标准，例如欧洲电信标准化协会（ETSI）制

定的推荐性标准。如果尚不存在统一标准，或者相关的统一标准被认定为存在不足，又或者需要解决特定的安全或基本权利问题，可以通过法案的方式设立共同规范。

（4）中国着眼人工智能产业全链路节点，有序平衡产业发展需求与安全标准建设

目前，我国人工智能安全风险领域标准研制工作正加速推进，在国家标准方面，主要涉及人工智能安全应用及个人隐私保护领域，智慧城市、自动驾驶等重点应用领域的标准研制力度逐渐加大。

2020年7月，国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部联合发布了《国家新一代人工智能标准体系建设指南》，其中人工智能安全/伦理标准被列为八个重点方向之一，包括人工智能领域的安全与隐私保护、伦理等部分，涵盖人工智能领域基础安全，数据、算法和模型，技术和系统，安全管理和服务，安全测试评估，产品和应用等相关标准，以及涉及传统道德和法律秩序的伦理标准，提出重点研究医疗、交通、应急救援等领域伦理标准，保障人工智能产业健康有序发展。

同时，在诸多应用领域，我国现已出台或立项一系列行业和团体标准：人工智能安全风险管理基础共性标准方面，各级标准化组织已立项相关标准或研究报告；人工智能安全应用方面，2019年，全国信息安全标准化技术委员会（TC260）立项《信息安全技术 人工智能应用安全指南》标准研究项目，内容包括研究人工智能的安全属性和原则、安全风险、安全管理及在需求、设计、开发训练、验证评估、运行等阶段的安全工程实践指南；数据安全方面，主要围绕大数据安全、数据交易服务安全等领域出台能力评估和安全管理标准，现已发布GB/T 35274 《信息安全技术 大数据服务安全能力要求》、GB/T 37932 《信息安全技术 数据交易服务安全要求》等标准；对于算法模型的评估，已立项的国家标准《机器学习算法安全评估规范（征求意见稿）》给出了机器学习算法安全评估指标、评估流程以及在需求、设计、开发训练等全流程阶段的算法安全评估规范。

表2 欧洲电信标准化协会设立的人工智能安全标准

项目	描述
SAI 001 AI Threat Ontology	定义什么是AI威胁
SAI 002 Data Supply Chain Report	保障AI全链路安全的现有技术
SAI 003 Security Testing of AI	AI测试指导
SAI 004 Securing AI Problem Statement	描述AI面临的安全挑战
SAI 005Mitigation Strategy Report	总结现有的AI风险消解方法
SAI 006 The role of hardware in security of AI	识别硬件在AI安全中的作用
SAI 007 Explicability and transparency of AI processing	研究AI的可解释性和透明性
SAI 008 Privacy aspects of AI/ML systems.	研究AI/机器学习系统中涉及的隐私问题
SAI 009 AI computing platform security framework	研究AI基础底座的安全功能与服务能力

资料来源：国家工业信息安全发展研究中心整理

2.3 人工智能算力基础设施安全保障目前主要参照平行领域标准，制定高质量可持续发展标准势在必行

由于人工智能算力基础设施属于近年来涌现出的新型实践模式探索，目前世界范围内尚无专门针对人工智能算力基础设施安全构建的直接实践。参照平行领域的安全标准能够满足基本维度的安全要求，针对人工智能算力基础设施各项具体特性指标的安全标准目前也已步入推进状态。

（1）人工智能算力基础设施建设标准主要对标关键信息基础设施、云服务和人工智能系统标准

根据人工智能算力基础设施帮助用户通过网络按需按量调用的云服务特性和大规模集成数据开展计算的实际特点，在规划建设过程中，服务供应商可参照国内外现有的云计算和数据中心安全标准，针对重点考量维度进行参照比对评估，能够满足人工智能算力基础设施的场地设备安全、软硬件可靠性、数据信息安全等多个维度的基本安全要求。下列表格列举了一些较具代表性的现有标准：

（2）面向人工智能算力基础设施特性指标的安全发展标准正加紧研判与建构

近年来，我国迎来了人工智能算力基础设施的建设热潮，各地纷纷筹划布局智算中心。为了在保证基本安全要求的前提下，进一步实现人工智能算力基础设施的高质量建设与运用，充分提升设施利用效率、避免重复建设，专门针对人工智能算力基础设施这一细分领域，以及算网融合传输能力、人工智能运算精度性能、适用任务场景、定价收费、兼容性与通用性能、电力能耗、不同算力网络间的互联互通等等具体人工智能算力基础设施特性研制安全标准，势在必行。

目前，人工智能算力基础设施特性相关新标准的编制立项工作，已经逐步得到了我国产学研各界的重视。例如，2022年2月，中国通信学会算网融合标准工作组组织召开标准评审会议，针对人工智能算力基础设施中的算网融合功能特性，提交《算网基础设施 总体能力要求》《算网基础设施 成熟度评价模型及关键指标》《算网基础设施 测试评估方法》《算网基础设施 网络互连》《算网基础设施

表3 我国云计算服务和数据中心安全规范标准

项目	标准名称
云计算服务相关安全标准	基于ISO/IEC 27002信息安全控制策略标准的ISO/IEC 27017云服务信息安全认证标准
	ISO/IEC TR 3445:2022云计算服务审计标准
	GB/T 34982-2017《云计算数据中心基本要求》
	GB/T 31168-2014《信息安全技术 云计算服务安全能力要求》
数据中心相关安全标准	YD/T 2585-2016《互联网数据中心安全防护检测要求》
	GB 51195-2016《互联网数据中心工程技术规范》
	ISO/IEC TS 22237 信息技术 数据中心设备与基础设施建设通用概念标准
	GB/T 51314-2018《数据中心基础设施运行维护标准》

资料来源：国家工业信息安全发展研究中心整理

安全要求》五项立项申请，为人工智能算力基础设施相关特性指标的标准建立、推动领域高质量可持续发展，迈出了坚实探索步伐。

（3）国际标准化组织着力推进现有安全评估标准体系与算力基建的深入融合发展

面向广义上的计算机和信息技术产品或系统安全评估工作，ISO组织构建了一个通用性的标准评估框架，推出了ISO15408信息技术安全评估通用标准（Common Criteria，CC），并推动成立了国际CC互认组织（CCRA）。在这一框架下，围绕IT产品内部资产保护、提升可靠性等不同产品形式的共同需求，要求产品开发者从资产出发，明确IT产品保护的资产，分析资产所有的安全问题，可以针对信息系统中某一具体硬件、软件、固件产品的安全提供评估。用户使用保护配置文件指定的安全功能要求和安全功能保证要求，技术供应商对其产品的安全属性进行声明，并委托测试实验室评估其产品以确定它们是否满足这些声明，并以EAL评估保障分级的形式，由一级至七级，呈现不同的安全性能评级。

在人工智能算力基础设施安全构建的实践当中，目前各服务供应商可参考CC安全标准框架，针对算力系统中所使用到的具体软件、硬件产品开展评估，从关键设备层面保障安全。例如，华为数据中心交换机和数字机房微模块控制器产品通过独立认证机构安全评估，获评CC认证中网络设备类别的最高EAL安全认证等级证书,表明两款设备已经达到了世界同类设备的前沿安全评级水准。将通过CC认证的设备运用于算力系统中，能够有效凸显系统的可靠性，因此在未来，更多企业将进一步关注产品的安全性能认证，共同构建人工智能算力基建领域的可信化生态环境。

三 技术工具研制方面，多主体发力人工智能安全管理能力提升，人工智能算力基础设施“助力安全”生态不断增强

为了实现人工智能安全治理原则的落地可用、降低安全治理难度，以一些国家和科技企业为代表的人工智能安全治理参与主体近年来推出了一批具有较好拓展性的技术评估和安全增强工具，有望与人工智能算力基础设施进行有机整合进而推广普及，未来也将借助人工智能算力基础设施提供的强大算力和新型服务模式实现进一步的效能优化和模式创新。

3.1 多个国家推出人工智能安全评估工具

国家作为推动人工智能安全治理的重要主体和牵引力量，应当着重在促进创新发展和维护应用安全两个方面之间不断探索平衡点，既要鼓励科技企业不断创新探索前沿应用，又要坚持安全治理底线，因此人工智能应用的安全水平评估类工具成为了各地区和国家重点探索的方向。

2020年7月，欧盟人工智能高级别专家组发布了正式版《可信人工智能评估清单》，围绕7项关键要素（实现人类自主和监管、确保技术稳健和安全、重视隐私和数据治理、注重透明性、注重多样性非歧视、实现环境友好和社会福祉、建立问责制）提出了可信人工智能的评估方法。

2022年6月，新加坡发布了人工智能治理测试框架和工具包“A.I. Verify”的最小可行产品(MVP)，成为全球首个国家层面发布的人工智能治理测试工具，该工具将测试和过程检查相融合，能够帮助企业建立与利益相关者之间的信任，旨在促进公众对人工智能技术的信任，同时支持人工智能技术的广泛使用。该工具有机融合了多种人工智能伦理原则框架，将11个关键的人工智能伦理原则合并在一起，分为透明度、决策过程、公平、管理与监督等5项支柱。其中11项原则具体包括透明度、可解释性、可重复性、安全性、安保性、稳健性、公平性、数据管理、问责制、人力资源及监督，以及包容性增长、社会与环境福祉。通过建设开箱即用的安全工具，有望打通人工智能治理框架原则之间的界限、提升新加坡在国际标准制定中的参与度。

基于相似的考量，加拿大政府也发布了“算法影响评估”这一风险评估工具，以帮助机构更好地预估自动决策系统带来的风险并制定相应的应对措施。该工具采用清单式和问卷式评估框架，通过设定10类与自动化决策系统业务流程、数据和设计相关的问题，以计算得分的方式自动确定算法的影响级别。评估指标重点强调数据质量问题检测流程、系统设计、开发、维护和改进的归责机制和措施以及规范检测数据偏差等技术防治措施。指标中风险概况部分包括四个主要问题，要求提供算法是否属于敏感行业、领域，是否产生重大影响等相关内容。指标中的风险应对措施部分，就数据质量、程序公平和隐私三方面内容共提出了30个具体问题。

3.2 企业积极推出人工智能安全相关技术和工具

着眼于协助人工智能安全治理的有效落地实施、促进人工智能良善运用，各国科技企业针对人工智能应用的开发和运维环节，积极研制一系列的安全增强赋能工具。

例如，微软致力于推动人工智能在人机交互和协作、公平性、可理解性和透明度、隐私、可靠性和安全性等方面的实践，设计了人工智能公平性检查工具、数据集数据表工具、Counterfit检测工具、InterpreteML工具等。其中，公平性检查工具帮助团队在人工智能生命周期每个阶段的决策进行

检查，从而帮助在部署之前预测和改善公平性问题。数据集数据表工具建议每个数据集都附有一个数据表，记录有关其创建、关键特征和限制的相关信息。Counterfit通过命令行测试给定的人工智能系统在作为开源平台时的稳定性和安全性、InterpreteML作为python工具包/库，集成了一系列可解释人工智能的前沿方法，帮助用户训练一个可解释的模型或解释黑箱模型。华为开发了一系列AI模型安全性增强工具：鲁棒性评测工具提供对抗样本检测、防御及攻防评测，可检测模型鲁棒性及威胁程度；用户隐私保护工具通过安全聚合训练、SignDS差分隐私训练技术保护用户隐私；数据漂移检测工具提供时序和图像数据的概念漂移检测，保障模型可用性。我国的人工智能安全解决方案提供商瑞莱智慧则聚焦通过工具构建，保证应用合规、隐私数据安全，围绕算法可靠、数据可用、应用可控三大方向打造多款人工智能基础安全平台：在算法安全方面，重点研究人工智能对抗攻防技术，提供模型安全性测评及防御加固的端到端解决方案；在数据安全方面，打造数据安全共享基础平台，实现了机器学习和分布式联邦学习生态的统一、加密算法的高效率优化、细粒度展示执行流程的安全评估验证能力；在人工智能应用治理领域，推出深度合成内容检测平台DeepReal与深度合成内容制作平台，重点反制“AI换脸”等深度伪造技术滥用现象。

针对发展中的人工智能算力基础设施，算力服务提供商也着眼服务用户的全业务流程，不断提升服务的安全性能。例如，围绕多用户共享的算力环境下模型厂商核心知识产权数据与模型的安全，华为基于可信硬件提供的密钥安全存储和授权技术，在人工智能算力基础设施运行环境中部署了模型与数据高效保护方案：模型厂商作为训练数据、模型的所有者，在其本地对训练数据和模型执行加密，并负责保护加密密钥，直接与AI计算中的可信环境建立信任关系，通过安全通道将密钥安全地注册到计算中心的硬件可信环境中，完成对数据和模型的加密传输；在此基础上计算中心进一步为训练或推理容器镜像提供完整性保护，并通过密钥动态授权容器运行时的身份认证，实现权限最小化，从而提升安全防护水平、降低受攻击的风险。

第四章
人工智能算力基础设施安全发展建议

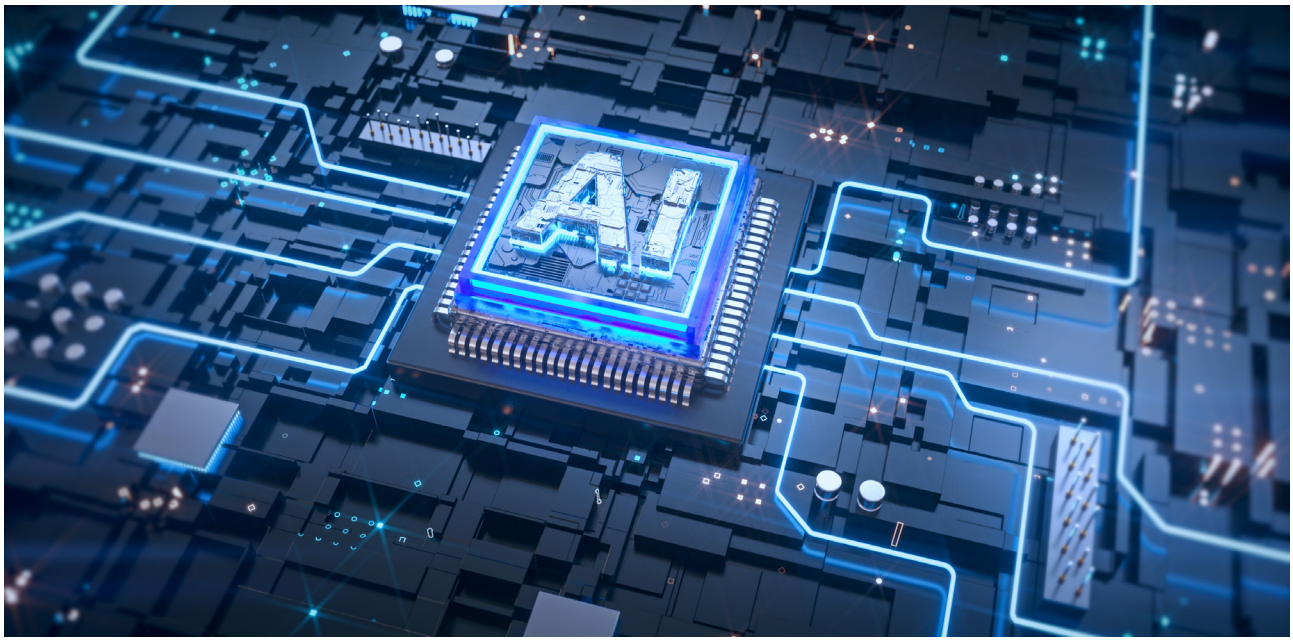
在政策推动与人工智能发展需求的牵引下，人工智能算力基础设施迅速落地发展，算力网络逐渐形成。与此同时，人工智能算力基础设施安全问题已引起多方关注，各类主体虽然已从政策、标准、技术等方面加强对关键基础设施安全和人工智能安全的保障，但专门针对人工智能算力基础设施安全发展的相关措施仍在探索当中，应从完善顶层设计、加快标准研制、加强技术攻关、建立管理制度等方面入手，更好应对和解决人工智能算力基础设施面临的安全问题，打造安全的人工智能算力底座，夯实我国人工智能产业健康发展的基础。

一 完善顶层设计，重视人工智能算力基础设施安全的新需求与新挑战

在关键基础设施建设中，对物理安全、网络安全等算力基础设施安全风险的广泛关注由来已久。随着人工智能的快速发展与应用，数据投毒、对抗样本、模型窃取等更多新兴安全风险日益突出，人工智能算力基础设施安全面临更多新挑战。同时，随着国家“东数西算”工程全面启动，构建算力网络成为趋势，然而算力网络融合、互连、灵活的特点也意味着对算力基础设施的网络、数据、应用等更高的安全要求。需要从战略规划、政策制定、资金引导等方面加强统筹布局，推动出台和实施相关政策，保障人工智能算力基础设施安全。一是针对人工智能算力基础设施面临的网络安全、物理安全等基础性挑战，应探索制定更符合人工智能和算力网络特点的相关配套政策，做好行业政策体系与《中华人民共和国网络安全法》《关键信息基础设施安全保护条例》等法律法规的衔接和落实。二是围绕人工智能面临的数据模型窃取、数据投毒、后门攻击、对抗样本攻击等新型安全问题开展前瞻研究，制定专门政策，要求从人工智能算力基础设施从设计之初就提前谋划和考量算法、数据等各环节的安全风险，确保设施安全稳定运行。

二 加快标准研制，构建基础设施安全与人工智能安全相融合的标准体系

人工智能算法需依赖算力和数据进行训练，因此，算法、算力、数据的安全问题难以分割，算力基础设施与人工智能的安全问题相互交织，一旦应对不当，可能会叠加放大。然而，目前针对人工智能算力基础设施的安全标准尚不统一，需要统筹考虑，实现协同联防，构筑具有针对性的一体化、全链路的标准体系。一是亟需制定相关标准并加快推动标准落地，明确人工智能算力基础设施安全的基准指标，使人工智能算力基础设施在能力建设、安全要求等方面满足一定准则，有效保障人工智能算法训练、运行过程中的环境安全；二是加快建设人工智能算力基础设施保障运行安全和助力安全合规等方面的相关标准，帮助提升人工智能算法安全性，推动形成行业健康发展的良性循环。



三 加强技术攻关，推动人工智能安全工具与人工智能算力基础设施集成

一是要加快安全检测、安全评估等相关技术工具研发。目前国外已围绕人工智能安全问题推出了多项技术工具，我们应围绕数据安全、算法公平、隐私保护等问题突出的领域，鼓励龙头企业大力开发安全技术工具，以国家科技重大专项、重大工程为依托，加大资源投入和创新要素整合，加快推动人工智能算力基础设施安全保障及安全工具技术的创新和演进，加大对新兴安全技术研发应用和成果转化的支持力度。二是要推动相关技术工具嵌入和集成到人工智能算力基础设施中，鼓励基础设施企业和算法企业加强合作，通过提供安全的算力基础设施，为算法开发者提供安全、可信的算力环境，通过集成相关技术工具支持模型、数据和应用的安全，有效降低企业部署和应用安全人工智能系统的门槛。

四 建立管理制度，形成管理手段与技术手段相结合的安全发展良好氛围

人工智能算力基础设施面临的安全风险多种多样，仅通过技术手段难以覆盖众多风险种类，还需要通过管理手段，将安全策略和安全控制融入到人工智能算力基础设施设计、建设、运行、维护的生命周期各阶段。一是要完善安全管理规章体系，指导人工智能算力基础设施建设与运营方落实安全主体责任，开展安全防护检查与风险评估，及时排查各类安全问题隐患。二是要推动人工智能算力基础设施上的管理手段建设，提供人工智能安全风险管理能力评估工具，帮助算法企业建立自上而下的产品安全组织架构，将安全理念融入人工智能产品研发设计、测试开发、部署上线、运行维护、退役下线等全生命周。

参考文献

[1] 郭鑫.信息安全等级保护测评与整改指导手册[M]. 机械工业出版社, 2021.

[2] 陈宇飞,沈超,王骞,等.人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10):16.

[3] 邱勤,徐天妮,于乐,等.算力网络安全架构与数据安全治理技术[J]. 信息安全研究, 2022, 8(4):11.

[4] GB/T 21052-2007,信息安全技术 信息系统物理安全技术要求[S].

[5] 汤志伟,李昱璇,张龙鹏.中美贸易摩擦背景下"卡脖子"技术识别方法与突破路径——以电子信息产业为例[J]. 科技进步与对策, 2021, 38(1):9.

[6] 国家工业信息安全发展研究中心.《智慧城市人工智能计算平台白皮书》,2021.

[7] 张琳琳,王腾.人工智能安全国际标准化进展研究[J].信息通信技术与政策,2021(11):73-78.

[8] 孔勇,韩继登,王义华.强制关键基础设施网络事件报告加强勒索软件攻击应对措施——美国《2022年关键基础设施网络事件报告法案》解读[J].中国信息化,2022(04):41-46.

[9] 宋钊,孙骞.人工智能背景下全球关键信息基础设施安全挑战与对策[J].信息安全与通信保密,2022(06):94-101.

[10] 余晓晖.开启关键信息基础设施安全保护新阶段[J].网络传播,2021(08):32-35.

[11] 高原,吕欣,李阳,穆琳,韩晓露,鲍旭华.国家关键信息基础设施系统安全防护研究综述[J].信息安全研究,2020,6(01):14-24.

[12] 孔勇,范佳雪.美国《关键基础设施识别、优先排序和保护》解读[J].中国信息化,2022(08):38-41.

[13] 杨诗雨,桂畅旎.美国网络安全和基础设施安全局（CISA）网络安全漏洞治理政策分析[J].中国信息安全,2022(06):34-39.

[14] 姜红德.面对工信安全，欧美国家如何应对？[J].中国信息化,2019(10):32-34.

[15] 姜波,程光.德国IT信息安全标签制度[J].工业信息安全,2022(05):6-11.

[16] Center for Data Innovation. U.S. AI Policy Report Card. 2022.

[17] 石建兵.欧洲网络与信息安全局《关键信息基础设施领域的物联网安全基线指南》[J].信息安全与通信保密,2018(01):80-95.

[18] 冯燕春.加快构建国家关键信息基础设施安全保障体系[J].中国信息安全,2016(11):42-46.

国家工业信息安全发展研究中心

地址：北京市石景山区鲁谷路35号
电话：010-88686077
邮编：100040

免责申明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何承诺，作者不对您在本文档基础上做出的任何行为承担责任。作者可能不经通知修改上述信息，恕不另行通知。