TC260-TR-004-2026

网络安全标准化技术研究报告

——工业具身智能安全标准化研究

v1.0-202603

全国网络安全标准化技术委员会秘书处

全国网络安全标准化技术委员会新技术安全标准特别工作组（SWG-ETS）

2026 年 3 月

本文档可从以下网址获得：
www.tc260.org.cn/

前　言

《网络安全标准化技术研究报告》（以下简称《技术报告》）是全国网络安全标准化技术委员会（以下简称“网安标委”）秘书处组织编制和发布的技术研究类文件。本文件立足网络安全和新兴技术应用领域前沿动态，通过系统的技术研究、产业调研、标准分析与综合研判，梳理关键领域网络安全风险与挑战，提出标准化发展趋势及相关工作实施建议，为网络安全国家标准制修订与网络安全保障实施提供前瞻性的技术参考与决策支撑。

声　明

本《技术报告》版权属于网安标委秘书处，未经秘书处书面授权，不得以任何方式抄袭、翻译《技术报告》的任何部分。凡转载或引用本《技术报告》的观点、数据，请注明“来源：全国网络安全标准化技术委员会秘书处”。

技术支持单位

本《技术报告》得到中国电子技术标准化研究院、东北大学、中国网络安全审查认证和市场监管大数据中心、中关村实验室等单位的技术支持。

主要编写人员：周睿康、蔡一鸣、夏冀、李渝哲、武卓、伍扬、刘栋、马梦娜。

一、工业具身智能发展现状

（一）具身智能、工业具身智能与工业机器人

具身智能是将人工智能融入机器人等物理实体，赋予它们感知、学习和与环境动态交互的能力。具身智能通过赋予机器“大脑”，使其具备感知、决策与执行的全栈能力，未来将在工业、医疗、服务等领域广泛应用。

工业具身智能是指具身智能技术在工业领域的深度应用，其核心在于将人工智能与工业机器人深度融合，通过物理实体与制造环境的实时交互，实现感知、决策与执行闭环，从而提升工业生产的智能化水平。

工业具身智能是具身智能在工业领域的“特化版本”，既承接其核心框架，又针对工业需求进行了优化和重构。工业具身智能和具身智能的区别与联系主要表现为：

a）具身智能广泛应用于家庭、驾驶、医疗等领域，而工业具身智能主要应用于制造、物流等工业场景；

b）具身智能的应用环境体现为开放、动态，而工业具身智能的应用环境主要表现为半结构化和高标准化；

c）具身智能的应用核心需求为交互的自然性，而工业具身智能的应用核心需求为可靠性和安全性。

工业机器人是一种在工业自动化中使用的、自动控制、可重复编程、多用途的操作机，可对三个或更多轴进行编程，其可以是固定式

或移动式，用于工业自动化应用中。相比于工业机器人，工业具身智能额外增加了人工智能属性。工业具身智能和工业机器人的区别与联系对比如下表所示。

表 1 工业具身智能和工业机器人对比

概念	工业机器人	工业具身智能
核心本质	自动化设备	工业领域的具身智能体
智能水平	预设程序、固定逻辑	自主感知、决策与学习
感知与环境交互	依赖精密的外部环境构建（如轨道、围栏），感知能力弱，与环境隔离。	多模态主动感知（视觉、力觉等），能与环境安全、自适应地交互。
任务灵活性	擅长重复性、结构化任务，换产线需大量重新编程和部署。	能够处理非结构化、不确定的任务，能快速适应产线和小批量生产的变化。
与人的关系	通常被围栏隔离，或通过技术手段实现有限的安全协作。	工业具身智能可以理解人的意图并主动配合。
典型应用	焊接、喷涂、装配（固定工位）等。	复杂装配、柔性抓取、产线维护、物流分拣（非预设环境）等。

（二）工业具身智能核心技术特征

工业具身智能的本质是“具身认知+自主控制+安全执行”的统一体，是实现智能制造高级阶段（如智能产线、工业大脑、人机协同系统等）的技术基础。工业具身智能主要具备以下关键技术特征：

a）在多模态感知方面，工业具身智能通过融合视觉、听觉、力觉、温度、振动等多种感知模态，实现对复杂工业环境的高精度、高鲁棒性感知。传感器与边缘智能芯片的深度耦合，使智能体具备实时

状态监测与环境建模能力。

b）在认知与决策协同方面，在认知层面，具身智能引入大模型、知识图谱与元学习等技术，使其具有情境理解、自我适应和经验迁移能力。决策层强调强化学习、多目标优化与博弈推理等方法，用于实现复杂任务的动态规划与最优策略制定。

c）在执行与控制自适应方面，依托柔性机器人、智能执行器与精密控制算法，工业具身智能能够在高动态、强扰动的物理环境中，实现安全、高效的操作行为。控制系统广泛引入模型预测控制（MPC）、学习控制（LBC）与自适应鲁棒控制，支撑执行层的动态调整能力。

d）在系统级实时性与闭环自治方面，具身智能强调感知-认知-决策-执行的高频闭环协同，依托边云协同计算架构，实现毫秒级的响应与策略刷新，满足高时效、高精度的工业运行需求。

（三）工业具身智能应用发展趋势

当前，工业具身智能正逐步从试验性部署向规模化应用过渡，其发展趋势主要体现在以下几个方面：

a）从固定场景向多场景泛化迁移：初期工业具身系统主要应用于结构化作业场景，如流水线质检、协作装配等，现已逐步向非结构环境（如仓储物流、柔性制造）扩展，强调系统对环境变化的泛化与适应能力。

b）从单体智能向群体协同演进：具身智能单体设备正发展为多智能体系统（MAS），通过边缘协同控制与任务分配机制，实现工业

机器人群体间的自组织、自协调与自优化。

c）从辅助操作向自主决策演进：越来越多的工业具身系统已从“人指导+机器执行”模式过渡到“机器自主决策+人监督”模式，特别是在极端环境作业、危险任务替代等场景中展现出巨大潜力。

d）标准化与安全性成为产业推进关键：随着工业具身智能系统复杂度提升，如何建立统一的接口标准、数据协议、安全等级评价体系成为制约其大规模落地的重要因素。

综上，工业具身智能将在未来制造系统的智能化、自主化与安全化升级中扮演核心角色，成为工业 4.0 迈向工业 5.0 的关键支撑技术。

（四）工业具身智能安全特性

工业具身智能系统作为高度集成感知、认知、决策与执行功能的智能体，其安全性评估需突破传统工业控制系统和信息系统的边界，构建具身层、认知层、控制层与网络层多维度协同的安全评估体系。本节从安全属性视角出发，对工业具身智能的关键安全特性进行系统评估。

a）本体安全（Physical Safety）

本体安全是工业具身智能系统的第一道防线，主要指系统在与物理世界交互过程中，所展现出的结构可靠性、执行稳定性与人员安全防护能力。由于具身智能体往往直接作用于设备、材料乃至人员，必须通过机械限力、物理隔离、急停机制、运动范围限制等方式确保不会在执行过程中引发冲撞、夹击、坠落等物理伤害。此外，系统应具

备完备的故障自诊断和容错控制机制，能够在异常情况下迅速响应、停机或切换至安全模式，从而保障工业现场作业人员的生命安全与财产安全。

b）网络安全（Cyber Security）

网络安全防护的核心目标是保障工业具身智能系统在运行过程中所采集、传输与处理的数据的完整性、机密性与可用性，防止其被篡改、泄露或伪造，覆盖从传感器到控制中心的全链路信息保护。鉴于该类系统普遍依赖网络通信、边缘计算与远程协同，存在感知数据被恶意注入、控制指令遭中间人劫持、模型参数被后门植入等风险，直接危害工业具身智能系统感知准确性、决策可靠性与行为安全性。因此，必须在工业具身智能系统架构层引入数据加密、安全身份认证、访问控制与可信计算机制，从而实现从硬件、通信、软件到算法的多重网络及数据安全防护。

c）控制安全（Control Robustness）

控制安全强调具身智能系统在执行动态任务过程中对扰动、干扰及攻击的鲁棒性与稳定性保障。由于系统需在复杂环境中进行精确控制操作，若控制环节遭遇干扰或模型失配，可能导致轨迹偏移、动作错误乃至任务失控。因此，必须构建具备冗余控制、模型自适应和状态反馈闭环机制的控制系统，确保在遭遇突发扰动、设备老化或外部攻击时仍能保持任务执行的连续性与安全性，从而实现系统级的容错控制与稳态恢复。

d）认知安全（Cognitive Trustworthiness）

认知安全主要指具身智能系统中的学习模型、感知算法与决策逻辑在长期自主运行中保持可信、稳定、不被操控的能力。具身智能系统往往嵌入深度学习、大模型、强化学习等复杂算法结构，在面对开放环境与不确定任务时，可能因样本污染、模型过拟合或外部诱导而生成不合理的策略行为。因此，系统必须具备算法行为可解释、输入输出监测、模型版本审计与抗对抗攻击能力，确保认知层的推理逻辑具备可信性、透明性与安全边界，防止认知偏差或被攻击引发系统整体风险。

二、工业具身智能安全风险分析

（一）工业信息物理系统安全风险

聚焦工业信息物理系统（CPS）在网络通信层面面临的潜在威胁，包括数据篡改、拒绝服务攻击（DoS）与通信链路的脆弱性，分析具身智能设备在开放环境下的网络安全防护需求。工业具身智能深度依赖的网络化控制架构，正成为系统性风险的放大器。通信协议标准化滞后导致设备间数据接口暴露攻击面，恶意攻击者可利用协议漏洞实施中间人攻击，篡改机械臂运动轨迹参数，造成加工精度失控或设备物理碰撞。更隐蔽的风险在于时间敏感网络（TSN）的确定性传输机制可能被劫持，攻击者通过精准时序扰动破坏多机协同节拍，触发生产节拍链式崩溃。此类风险直接威胁产线物理安全，可能引发千万级设备损毁事故。

（二）工业数据安全与隐私泄露风险

探讨工业具身智能在数据采集、传输与存储过程中可能暴露的安全漏洞，尤其是敏感工业数据和用户隐私面临的非法访问、泄露和滥用等风险。多源异构数据的聚合分析加剧了工业知识资产流失风险。高价值工艺参数（如注塑成型温度曲线）在边缘计算节点的临时存储过程中，可能因内存保护机制缺陷遭侧信道攻击窃取，导致企业核心竞争力被逆向工程破解。更具破坏性的是，工业视觉系统采集的员工操作行为数据若遭泄露，不仅违反个人信息保护法规，更可能被用于社会工程学攻击，通过分析操作习惯弱点实施定制化网络渗透。

（三）工业人工智能算法模型安全性与可解释性

分析工业具身智能依赖的 AI 算法在模型决策过程中的不透明性问题，重点关注模型攻击（如对抗样本）、训练数据污染以及算法不可解释性所带来的安全隐患。黑箱化决策机制正在制造不可预知的失控隐患。深度强化学习模型在动态环境中的自适应行为缺乏可解释性支撑，例如 AGV 路径规划算法可能因奖励函数设计偏差，在避障逻辑中隐含选择撞击货架的次优解。更危险的是，对抗样本攻击可针对性欺骗工业检测模型，使质检系统对关键缺陷视而不见，导致批次性质量问题流入市场。这种算法层面的不可控性，使得工业智能系统犹如“定时炸弹”，随时可能触发质量安全事件。

（四）人机协作过程中系统失控风险

识别在工业现场中人机共融交互过程中可能出现的行为偏差、操

作冲突与误判等情况，指出其对生产安全性与操作可靠性的潜在影响。认知不对称导致的人机互信危机正在酝酿新的伤亡风险。当协作机器人基于不可解释的决策逻辑突发异常加速时，操作人员因无法预判其行为轨迹，可能陷入物理闪避困境。在汽车总装场景中，这种风险尤为致命——机械臂的不可预测摆动可能将工作人员挤压于工装夹具之间。现有安全防护机制依赖预设的电子围栏与急停按钮，但面对智能设备自主决策产生的突发性危险动作，传统手段存在毫秒级响应延迟的安全真空。

（五）系统性能评估不足与鲁棒性缺陷

评估工业具身智能系统在复杂、动态环境中的性能稳定性与抗扰能力不足问题，强调缺乏标准化评估指标与验证机制所带来的安全隐患。复杂环境下的性能退化可能引发灾难性失效。激光导航系统在粉尘浓度超标时产生的建图偏差，若未纳入出厂检测标准，将导致 AMR 集群在雾霾天气集体“失明”，引发全厂区物流瘫痪。更严峻的是，多模态感知系统的交叉验证机制存在设计漏洞，当视觉系统因反光干扰失效时，力觉反馈可能无法及时接管控制权，造成精密装配环节的批量性零件损毁。这种评估盲区使得工业智能设备成为“薛定谔的安全体”。

（六）故障与恶意攻击共存的复合安全威胁

剖析具身智能系统在实际运行中可能同时面临自然故障与恶意攻击的复合型安全挑战，探讨系统在双重威胁下的韧性设计与防御策

略。自然劣化与人为破坏的叠加效应正在突破防御体系极限。减速器齿轮磨损产生的振动信号异常，可能被攻击者伪装成正常工况特征，诱导预测性维护系统延迟报警，最终导致产线紧急停机与设备永久性损伤。此类复合攻击充分利用系统固有弱点，使得传统基于单一威胁假设的安全设计彻底失效，可能造成指数级放大的损失——单个传感器的协同故障可能通过智能调度算法演变为全厂区生产大瘫痪。

三、工业具身智能安全政策与标准现状

目前，虽然尚未有专门针对“工业具身智能”的安全政策出台，但各国政府已在工业机器人、工业人工智能、工业互联网等相关领域发布了多项政策和标准，涵盖了安全治理、标准体系建设、网络安全防护等方面。

（一）国内外工业具身智能安全战略与政策法规

a）联合国推动人工智能伦理安全框架落地

联合国教科文组织于 2021 年 11 月发布《人工智能伦理问题建议书》，其提出的伦理原则（如透明度、责任性、人类监督）为人工智能安全治理提供了核心方向。建议书强调“人类中心主义”原则，要求协作机器人等典型工业具身智能系统需通过物理安全设计与算法伦理审查双重保障，确保在焊接、搬运等高危任务中优先保护人类操作员的生命安全。2023 年 3 月 31 日，该组织号召各国立即执行《人工智能伦理问题建议书》。

b）欧盟加强风险分级监管

欧盟委员会于2021年4月提出《欧盟人工智能法案》草案，经欧洲议会、成员国及委员会三方协商，于2023年12月达成协议，2024年3月欧洲议会正式通过，2024年8月生效。法案规则分阶段实施，其中禁止类条款于2025年2月生效，高风险系统合规要求于2026年8月全面执行。该法案旨在建立“以风险为基础”的监管体系，防范AI对民主、人权、环境等领域的威胁，同时推动可信赖AI技术发展，确立欧盟在全球AI治理中的标准制定权。该法案明确禁止了多种有害的AI实践。例如，严禁利用潜意识或操纵性技术影响人类行为，这种技术可能会通过不易察觉的方式干扰人们的自主决策；基于敏感特征进行生物识别分类也被严格禁止，以保护公民的隐私和防止歧视；在公共场所使用实时远程生物识别技术同样被禁止，不过在少数执法场景下除外。《欧盟人工智能法案》所提出的高风险AI系统涵盖了作为产品安全组件或用于特定高风险领域的AI系统。此类系统的提供商需进行严格的合规评估，必须满足风险管理、数据治理等多项要求。在部分情况下，还需要开展基本权利影响评估。若系统符合欧盟协调标准，则被推定为合规，但在上市后仍需进行持续监测和整改。该法案适用于所有在欧盟市场提供AI服务的境外企业，迫使全球科技公司调整合规策略。该法案标志着欧盟从“事后纠错”转向“前瞻预防”的监管思维，其分级治理模式既避免了过度限制创新，又为高风险领域划定了明确红线，或将成为未来全球工业具身智能治理的基准框架。

c）美国采取标准引导原则，鼓励企业行业自律

美国在人工智能安全相关领域采取“技术自治优先、多方协同治理”的策略，通过标准引导与行业自律相结合的方式，构建覆盖物理操作安全、数据完整性、人机协同可靠性的治理体系。《国家人工智能倡议法案》于2021年1月1日正式颁布，强化和协调各联邦机构的人工智能研发活动，确保美国在全球人工智能技术领域保持领先地位。根据该法案，美国设立了国家人工智能倡议办公室，隶属于白宫科技政策办公室，负责监督和执行美国国家人工智能战略等任务。同时，法案规定成立国家人工智能咨询委员会，该机构的职责是向总统和倡议办公室提供关于人工智能产生的法律问题以及相关的责任和法律权利的相关情况，例如：倡议是否解决人工智能带来的伦理、法律、安全或其他社会问题；关于人工智能治理路径的检视；人工智能系统违反现有法律的责任以及平衡个体权利与人工智能创新的发展情况等。

美国工业具身智能相关安全政策以灵活性和创新包容性为核心，通过标准引导而非强制立法，鼓励企业将安全能力嵌入产品全生命周期。这种模式在保障工业机器人、智能产线等场景安全的同时，也为技术快速迭代预留了空间。

d）其他国家均构建各自工业具身智能战略

1）英国：伦理引领、风险适配

英国《国家人工智能战略（2021）》将人工智能列为关键技术领

域，提出以“可信 AI”原则指导物理操作安全与算法透明度。工业机器人需通过英国标准协会（BSI）的 PAS 440 伦理审查框架，确保人机协作场景下的力控阈值符合 BS 8628 功能安全标准（如冗余制动系统设计），同时要求决策逻辑可追溯（如基于 ROS 的路径规划日志）。国家网络安全中心（NCSC）发布的工业控制安全指南进一步规定，人工智能设备需部署实时入侵检测系统，利用数字孪生技术模拟异常行为（如传感器信号欺骗攻击），并强制验证固件签名（RSA-3072 加密）。

2023 年《人工智能监管政策白皮书》提出“基于场景的风险分级”模型，对人工智能实施差异化监管。高风险场景（如核电站巡检机器人）需通过 UKCA 认证，满足 BS 10128 数据完整性规范（如激光雷达点云数据区块链存证），而中低风险场景（如仓储 AGV）遵循英国机器人协会（BRA）的 BRC GS9 行业自律标准，强制披露安全设计文档（如 ISO 10218-1 合规声明）。《数据保护与数字信息法案（草案）》要求视觉/力觉数据在边缘端完成匿名化处理（如 SLAM 建图数据模糊化），禁止跨境传输原始操作日志，以应对工业敏感信息泄露风险。

2）加拿大：保障民众权益

加拿大在工业具身智能相关安全领域以保障民众权益为核心，通过《2022 年数字宪章实施法案》及配套的《人工智能与数据法》（AIDA）构建了独特的治理框架。法案将工业场景中的人机物理交互系统（如

协作机器人、高危环境巡检设备）明确列为高风险监管对象，要求企业开发部署时必须实现数据匿名化处理与算法公平性验证。工业机器人在采集视觉、力觉等数据时需通过边缘计算节点进行去标识化，例如 SLAM 建图数据需模糊化处理以避免泄露工厂布局细节，同时严格禁止存储工人面部特征等生物信息。为消除算法偏见，加拿大标准委员会（SCC）制定《CAN/CIOSC 101:2023 算法偏见测试标准》，强制企业验证 AGV 路径规划、任务分配等决策逻辑的公平性，确保不因性别、种族等因素产生歧视性结果，相关测试需包含至少 5000 组涵盖多元劳动力特征的仿真场景数据。

3）俄罗斯：国家安全驱动、技术主权优先

俄罗斯在人工智能安全领域采取“国家安全驱动、技术主权优先”的战略导向，通过顶层立法与产业政策联动，构建覆盖军事、工业场景的人工智能安全治理体系。2019 年总统令批准的《2030 年前国家人工智能发展战略》明确将工业机器人、自主化武器系统等具身智能技术列为“国家技术主权核心领域”，要求突破西方技术封锁，实现关键硬件（如伺服电机、激光雷达）与算法框架（如联邦学习、强化学习控制模型）的完全自主化。该战略特别强调工业场景中的人机安全协同，规定军工复合体企业必须通过 GOST R 59276-2020 工业机器人功能安全认证，确保核设施巡检机器人、无人战车等设备在极端电磁干扰、网络攻击下的操作可靠性（如抗干扰通信协议、三冗余制动系统）。

2020 年俄联邦政府颁布的《至 2024 年人工智能和机器人技术监管构想》进一步细化法律适配机制，提出“人-机-法三元协同”框架。在工业领域，强制要求协作机器人部署方遵守《工业自动化安全法》修正案，包括动态安全距离实时标定（依据 GOST R ISO 10218-1 标准）、数据主权本地化（工业传感器数据仅存储于境内云服务器）及控制指令双因子认证（生物特征+物理密钥）。例如，俄罗斯国家原子能公司（Rosatom）在核电站引入的“阿尔法”系列巡检机器人，其 SLAM 建图数据需通过俄罗斯联邦安全局（FSB）认证的“穹顶”加密系统处理，并植入自研的对抗样本检测模块（基于 Kaspersky 工业威胁情报库），以抵御针对路径规划算法的欺骗攻击。

4）新加坡：深化人工智能在经济和社会各领域的应用

新加坡在工业具身智能领域尚未发布专门的国家级政策，但已通过人工智能、机器人技术、网络安全等相关政策和项目，积极推动具身智能技术的发展和应用。2019 年 11 月，新加坡金融管理局（MAS）联合金融机构推出 Veritas 计划框架，率先在金融领域构建人工智能伦理评估体系，要求银行、保险机构对 AI 驱动的信贷评估、欺诈检测等系统进行“公平性-道德性-透明性”三重审查。例如，星展银行（DBS）基于 Veritas 框架优化其智能投顾算法，通过动态调整风险评估模型参数，确保不同种族客户群体的贷款批准率偏差小于 5%。2022 年 5 月，新加坡资讯通信媒体发展局（IMDA）与个人数据保护委员会（PDPC）共同发布 A.I.VERIFY 工具包，将治理范围扩展至工

业与公共领域。该工具包整合技术测试（如对抗样本鲁棒性验证）与程序审查（如数据采集合规性审计），要求工业机器人等具身智能设备在部署前提交可解释性报告（如基于 SHAP 值的机械臂路径规划逻辑可视化），并在医疗、交通等高风险场景中强制实施第三方安全认证（如符合 ISO 31000 风险管理标准的动态避障性能测试）。

5）日本：平衡技术创新与社会风险防控

日本《人工智能战略 2022》提出“技术赋能与社会韧性并重”的治理方针，围绕“以人为本、多样性、可持续”三大原则细化安全举措。在工业具身智能相关领域，政策重点聚焦人机协作安全与重大灾害响应。总务省（MIC）发布《工业 AI 数据治理指南》，规定工业机器人采集的视觉/力觉数据需在边缘端完成匿名化（如丰田工厂采用同态加密技术处理焊接机器人操作日志），并禁止跨境传输涉及生产流程敏感信息的数据包（如激光雷达扫描的工厂三维模型）。

e）我国高度重视，有效平衡发展和安全

我国在工业具身智能相关领域的政策布局日益完善，已将其纳入国家战略和地方发展重点，涵盖技术研发、产业集群建设、应用场景拓展等多个方面。我国在《国家新一代人工智能发展规划》中明确提出推动工业智能化转型，同时强化安全底线。《生成式人工智能服务管理办法（征求意见稿）》虽聚焦生成式 AI，但其数据安全、算法透明性等要求对工业具身智能具有普适性。2023 年《机器人+应用行动实施方案》强调工业机器人需符合功能安全与网络安全双重标准，

要求企业建立全生命周期安全管理体系。

《智能制造发展规划（2021-2035 年）》要求工业机器人、自动化装备需通过安全认证，保障人机协作场景下的物理安全和数据隐私。2025 年 3 月，具身智能首次被写入《政府工作报告》，与生物制造、量子科技、6G 等并列为未来产业，提出建立未来产业投入增长机制，培育具身智能等新兴产业。

北京市发布《北京具身智能科技创新三年行动计划（2024—2026 年）》，目标是到 2027 年突破百余项关键技术，产出不少于 10 项国际领先成果，推动万台机器人规模落地，培育千亿级产业集群。深圳市科技创新局印发《深圳市具身智能机器人技术创新与产业发展行动计划》，支持具身智能机器人企业建设制造工厂，推动新产品、新技术首次应用和产业化，鼓励高质量具身智能数据集开放共享。

（二）国内外工业具身智能安全标准体系

当前尚无直接以“工业具身智能（Industrial Embodied Intelligence）”命名的完整标准体系，但可以从工业机器人、工业人工智能、智能制造、具身智能、网络安全等关键子领域出发，梳理国内外已有的标准体系，为构建工业具身智能安全标准提供参考。

a）与工业具身智能相关的安全国际标准

与工业具身智能相关的安全标准体系在全球范围内呈现多层次、多维度的协同发展格局。国际标准化组织（ISO/IEC）与电气与电子工程师协会（IEEE）共同构建了基础性框架，聚焦全生命周期安全与

可信能力。

ISO/IEC JTC1 SC42 分技术委员会主导的系列标准中，ISO/IEC 38507:2022 明确了工业机器人的治理框架，要求数据主权管理（如跨境传感器数据流动规则）与硬件冗余设计（三电系统备份机制）相结合，同时将高危环境下的功能安全认证（如核电站巡检机器人需满足 ISO 13849 PL e 级标准）纳入强制要求。

在可信性方面，ISO/IEC TR 24028:2020 规范了机械臂运动轨迹规划逻辑的可视化解释，要求通过 SHAP 值验证动态避障性能，而 ISO/IEC 27090 草案则针对工业控制指令的安全性，规定 ROS 通信协议需集成端到端加密（TLS 1.3）与 OPC UA 安全架构。

隐私与功能安全的融合设计由 ISO/IEC TR 5469 推动，例如力觉传感器数据需在边缘节点完成同态加密处理，并同步符合 ISO 10218-1/2 的机械精度指标（±0.02mm）。IEEE 的标准体系则侧重伦理责任，IEEE P7007-2021 要求工业设备配备“安全操作数字护照”，记录故障事件与伦理审查结果，而 IEEE 2841-2022 规范了工业质检场景下深度学习模型的对抗鲁棒性测试，覆盖至少 20 类攻击场景（如 FGSM、PGD）。

b）欧洲发布多份工业机器人、网络安全等相关安全指南文件及标准需求

欧洲通过 ETSI 与 CEN/CENELEC 推动行业细分化创新。ETSI 发布的 GR SAI 004 提出工业机器人数据安全“三防”机制——防窃取

（AES-256 加密存储）、防篡改（基于区块链的数据存证）、防滥用（最小化采集工件尺寸数据），而 GR SAI 005 则针对协作机器人任务分配算法，要求通过 BSI EN 626-1 公平性测试（确保无性别、年龄偏见）。CEN-CLC/JTC 21 发布的 CEN/TS 17982-2024 整合了机械安全（EN ISO 12100）、网络安全（IEC 62443）与伦理要求，例如数字孪生模型需通过误差率＜0.5%的 V&V 验证，联邦学习架构需满足 GDPR 数据本地化约束。

c）美国关注可信任可解释的人工智能研究

美国以 NIST 为核心构建风险驱动的标准生态，其 AI RMF 1.0 框架提出人工智能“五层防护链”：硬件层冗余制动系统（UL 3300 认证）、算法层对抗鲁棒性测试（NIST IR 8269）、数据层 SLAM 点云脱敏（NIST SP 800-53）。企业自律标准同步推进，谷歌的《工业 AI 安全白皮书》要求焊接机器人嵌入动态功率限幅模块，微软的《负责任工业 AI 标准》则规定数字孪生系统需模拟 PLC 指令劫持攻击并制定应急手册。

d）我国提前布局，即将出台多项人工智能安全标准

为积极响应《全球人工智能治理倡议》，落实《人工智能安全治理框架》，全国网络安全标准化技术委员会秘书处组织开展人工智能安全标准体系研究工作，牵头组织撰写《人工智能安全标准体系（V1.0）》（征求意见稿）。人工智能安全标准体系由基础共性、安全管理、关键技术、测试评估、产品与应用等 5 个部分组成，体系框

架如下图所示。

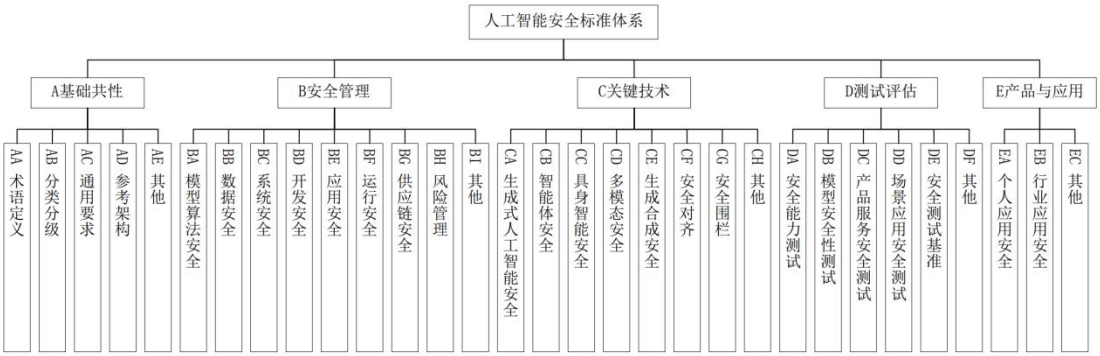

图 1 人工智能安全标准体系框架图

基础共性类标准是推动人工智能安全标准体系建设、落实《框架》各项措施的重要保障，是人工智能安全的基础性、总体性标准，包括术语定义、分类分级、通用要求、参考架构等研制方向。

安全管理类标准围绕《框架》中明确的模型算法安全、数据安全、系统安全三类内生安全风险，以及人工智能开发、应用、运行、维护等各环节面临的安全风险，提供了覆盖全过程全要素的安全管理标准，包括模型算法安全、数据安全、系统安全、开发安全、应用安全、运行安全、供应链安全、风险管理等研制方向。

关键技术类标准紧扣人工智能相关技术发展及应用情况，明确各项关键技术的安全保障要求，为人工智能技术健康发展保驾护航，包括生成式人工智能安全、智能体安全、具身智能安全、多模态安全、生成合成安全、安全对齐、安全围栏等研制方向。

测试评估类标准旨在以测试评估工作帮助提升人工智能安全水平，包括安全能力测试、模型安全性测试、产品服务安全测试、场景应用安全测试、安全测试基准等研制方向。

产品与应用类标准旨在保障人工智能在各行业、各领域的安全应用，包括个人应用安全、行业应用安全等研制方向。

目前，我国已发布人工智能安全强制性国家标准 1 项、推荐性国家标准 12 项、技术文件 1 项，部分与工业具身智能相关的标准如下：

GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》强制性标准规定了人工智能生成合成内容显式标识和隐式标识的种类、要素和格式，给出了人工智能生成合成内容标识方法。

GB/T 41871-2022《信息安全技术 汽车数据处理安全要求》标准规定了汽车数据处理者对汽车数据进行收集、传输等处理活动的通用安全要求、车外数据安全要求、座舱数据安全要求和管理安全要求。适用于汽车数据处理者开展汽车数据处理活动，适用于汽车的设计、生产、销售、使用和运维，也适用于主管监管部门和第三方评估机构等对汽车数据处理活动进行监督、管理和评估。

GB/T 42888-2023《信息安全技术 机器学习算法安全评估规范》标准规定了机器学习算法技术和服务的安全要求和评估方法，以及机器学习算法安全评估流程。适用于指导机器学习算法提供者保障机器学习算法生存周期安全以及开展机器学习算法安全评估，也可为监管评估提供参考。

GB/T 45654-2025《网络安全技术 生成式人工智能服务安全基本要求》标准提出面向我国境内公众提供生成式人工智能服务的基本安全要求，具体包括生成式人工智能服务相关的算法模型安全、训练数

据安全、数据标注安全、防范虚假信息、防止用户沉迷、防范歧视、保护用户隐私等方面内容。

GB/T 45674-2025《网络安全技术 生成式人工智能数据标注安全规范》标准针对生成式人工智能产品研制中人工标注环节的标注规则、标注人员培训、标注内容正确性等方面提出安全规范。

GB/T 45652-2025《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》标准针对生成式人工智能产品预训练和优化训练数据来源合法性，符合法律法规要求，不含侵犯知识产权内容，保护个人信息，保证真实性、准确性、客观性、多样性等方面提出安全规范。

GB/T 45958-2025《网络安全技术 人工智能计算平台安全框架》标准提出了人工智能计算平台安全框架以及相应的安全模块和机制，以保障用户数据与人工智能模型数据安全。

TC260-003《生成式人工智能服务安全基本要求》技术文件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，并给出了安全评估要求。适用于服务提供者开展安全评估、提高安全水平，也可为相关主管部门评判生成式人工智能服务安全水平提供参考。

TC260-PG-20233A《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》实践指南给出了生成式人工智能服务提供者对生成内容进行标识的方法。适用于生成式人工智能服务提供者利用生成式

人工智能技术向公众提供生成文本、图片、音频、视频等内容时对生成内容进行标识。

TC260-PG-20211A《网络安全标准实践指南—人工智能伦理安全风险防范指引》实践指南针对人工智能可能产生的伦理安全风险问题，给出了安全开展人工智能研究开发、设计制造、部署应用等相关活动的规范指引。适用于相关组织或个人开展人工智能研究开发、设计制造、部署应用等相关活动。

T/CESA 1193-2022《信息技术 人工智能 风险管理能力评估》团体标准覆盖与工业具身智能相关的工业人工智能算法与工业机器人双重安全，例如协作机器人需通过动态偏见检测和硬件失效模式分析，确保高危环境下的人机协作可靠性。

当前，我国正在推动的人工智能安全国家标准主要如下所示。

《网络安全技术 互联网信息服务深度合成安全规范》标准拟从互联网信息服务生命周期的信息生成、处理、发布、传播、存储、销毁等环节，以及技术算法生命周期的设计开发、验证测试、部署运行、维护升级、退役下线等五个阶段，对深度合成服务提供者和技术支持者提出开展互联网深度合成服务在安全方面的通用要求以及证实评估方法。

《网络安全技术 人工智能代码生成服务安全要求》标准拟针对利用生成式人工智能技术所构建的代码生成类互联网信息服务，提出文书签署、代码审查、 过程披露、风险提示等方面安全要求。

e）我国研制发布多项机器人安全标准

当前，我国机器人安全标准体系正在不断发展，安全内涵在持续扩展，从传统的机械电气安全，向功能安全和网络安全等维度延伸。

GB/T 20867.1-2024《机器人 安全要求应用规范 第1部分：工业机器人》是工业机器人安全的基础性标准，为设计、制造和集成提供了安全依据。

GB/T 45509-2025《工业机器人 动态稳定性试验方法》关注工业机器人在动态运动下的稳定性，从关键部件和性能层面保障整体应用安全。

GB/T 45501-2025《工业机器人 三维视觉引导系统 通用技术要求》规范了工业机器人三维视觉系统的技术架构和性能。

GB/T 45502-2025《服务机器人信息安全通用要求》是服务机器人领域的一项重要标准，规定了信息安全要求和测试方法，以防范数据泄露、网络攻击等风险。

除此之外，GB/T 39404-2020《工业机器人控制单元的信息安全通用要求》等标准正在开展修订工作。

四、工业具身智能安全标准需求分析

（一）网络安全防护标准需求

工业具身智能系统深度集成网络通信、边缘计算与工业控制技术，面临设备海量异构接入、通信协议多样、固件漏洞频发等严峻网络安全挑战。为此，亟需构建面向工业现场复杂环境的网络安全防护

标准体系，覆盖设备接入认证、通信传输安全、系统运行监控、漏洞管理及设备退网等全流程环节。针对工业现场多源异构设备互联场景，重点规范并细化设备身份鉴别、通信加密、入侵检测、漏洞修复等关键环节的网络安全防护要求。相关标准应明确规定设备间最小权限访问控制，并建立健全工业协议漏洞的威胁情报共享与协同处置机制要求，以体系化、规范化的标准全面提升工业具身智能在复杂网络环境下的安全防护能力。

建议研制标准：《工业具身智能网络安全防护基本要求》。

（二）数据安全与隐私保护标准需求

工业具身智能涉及核心生产工艺参数、关键设备运行状态及用户操作日志等数据。为有效管控数据安全风险，亟需构建涵盖数据采集、传输、存储、使用加工、共享、销毁等数据全生命周期安全标准体系，确立科学严谨的数据分类分级标准，配套制定强制的存储加密规范、可信的传输保护要求以及可追溯的共享审计机制。同时，应重点规范数据跨系统、跨域流动的安全评估方法与标准化流程，明确隐私计算技术在工业数据协同中的合规边界和技术要求，建立符合《数据安全法》及相关规定的工业数据合规性评估标准体系，确保在充分释放数据要素价值的同时，严格保障个人隐私与重要工业数据安全，支撑数据要素安全高效流通。

建议研制标准：《工业具身智能数据分类分级方法》《工业具身智能数据安全管理要求》。

（三）智能算法安全评估标准需求

工业控制算法的可靠性直接关系生产安全，需建立覆盖算法开发、部署、迭代的全生命周期评估标准。重点制定工业机器学习、预测性维护等算法的可解释性评价指标，研究算法的安全容错阈值与失效保护机制。规范工业领域专用智能算法的偏差检测与修正标准，确保算法决策符合物理约束与安全规程。

建议研制标准：《工业具身智能算法安全要求》《工业具身智能模型安全要求》。

（四）用户行为规范与操作安全标准需求

工业具身智能系统通常涉及人类操作员与工业机器人等智能设备的紧密互动，尤其是在制造业等高危环境中。需建立人机协同场景下的操作权限动态分级标准，实现操作行为的虚拟预演与风险预测，明确高危操作的电子围栏机制。推动工业智能化进入安全可控的新阶段。

建议研制标准：《工业具身智能交互安全要求》。

（五）安全性能指标与度量体系构建

随着工业具身智能安全治理体系建设的推进，安全属性定义在行业实践中逐步形成初步共识。国际标准及技术文件虽已对部分安全属性作出定义性描述，但不同框架间仍存在内涵界定与度量维度的差异性。明确系统安全评价内容，以及其量化度量方法，将成为工业具身智能发展的关键支撑。

建议复用工业机器人、关键工控设备相关标准。

（六）安全测试验证与合规性评估标准需求

工业具身智能通常会遭遇极端工况、网络攻击及人机协同中等安全风险。研究工业具身智能的安全验证理论与方法，指定不同安全性要求下的安全测试结果多级评估，有利于促进安全隐患提前防范，提升工业具身智能安全的信任度，降低法律风险。

建议研制标准：《工业具身智能可信评估指标》《工业具身智能算法安全评估规范》。

五、工业具身智能安全标准化工作建议

下一步，以工业具身智能安全标准化工作为抓手，扎实落实相关政策文件，积极推动标准体系构建、重点标准研制和标准宣贯推广等工作，促进工业具身智能安全防护能力提升。

（一）构建工业具身智能安全标准体系

针对工业具身智能安全风险覆盖面广、特征复杂等特点，系统性构建以网络安全与数据安全标准为核心基础，兼具跨领域、跨行业特性的工业具身智能安全标准体系，指导研制工业具身智能安全防护、测试等标准，满足石化、冶金、钢铁、电力等各工业行业在应用具身智能系统中的标准化需求。

（二）加强关键技术研究与重点标准研制

面向工业具身智能发展中的安全问题，加强网络安全、数据安全以及人机协作安全等方向关键技术研究，重点推动工业具身智能网络

安全及数据安全等相关标准的立项与研制工作。同时，加快推进《网络安全技术 工业控制系统网络安全防护能力成熟度模型》等在研标准研制发布进程，为产业安全发展提供标准化支撑。

（三）深入开展重点标准宣贯推广工作

依托全国网络安全标准化技术委员会、中国网络安全产业联盟等平台，在全国范围内分片区、分行业开展工业具身智能重点标准宣贯工作，面向各工业行业开展标准应用试点方案遴选工作，树立一批优秀典型案例，向全行业进行推广。