

《智能机器人伦理风险管理指南》编制说明

一、项目背景

（一）产业应用背景：智能机器人规模化普及，人机交互、多机协同双重交互伦理风险同步凸显

智能机器人集成 AI 感知、推理与自主决策能力，现已广泛应用于制造、养老、医疗、政务、金融、仓储等场景，在提质增效、补齐民生服务短板、推动产业数字化转型等方面释放巨大产业价值与社会效益。智能机器人涵盖具身机器人与无物理实体的机器系统（如数智员工、智能客服等）两类形态，兼具工具属性与自主赋权特征，形成人机交互、多机协同两类差异化交互模式，分别带来算法歧视、隐私侵害，以及任务失衡、故障级联、群体失控等伦理风险。

人机交互维度：机器人与人类近距离、长时间进行物理、情感、信息交互常态化，风险直接作用于自然人，社会关注度高：训练样本失衡引发年龄、外貌、地域等算法歧视；摄像、语音、雷达设备无限制采集居家、人脸等生理敏感信息，造成隐私泄露；深度学习算法黑箱导致自主决策不可追溯，伤人、财产损失事故权责划分困难；陪护类机器人过度拟人化，诱发老人、儿童情感依赖、现实社交弱化；设备底层系统存在安全漏洞，易被篡改操控，冲击公共安全秩序。

多机协同维度：物流、智造、智慧城市等场景批量部署多机器人协同系统，机器调度、跨设备数据互通、任务竞标、路径竞争等场景持续扩容，产生个体设备管控模式无法覆盖的次生风险：调度算法缺少公平约束，出现高风险任务集中分配；机器隐

性通信形成任务合谋，破坏市场公平；单台故障、恶意指令沿协同网络级联扩散，引发大范围服务瘫痪；大规模群体涌现行为挤占公共空间；跨厂商设备数据互通未脱敏，造成隐私大范围跨系统流转。此类多机协同风险隐蔽性强、传导速度快、危害具备放大效应，传统仅针对个体设备的安全、伦理管控体系存在天然短板。

基于两类交互风险的生成逻辑、识别方法、处置路径相对不同，本文件创新划分人机交互为核心、多机协同为补充双维度治理框架，同步适配多场景应用模式，实现各类交互风险精准管控、闭环处置。

（二）政策发展背景：国家、省市顶层文件明确要求加快机器人细分伦理标准供给

国家先后出台系列顶层政策、法规，对智能机器人伦理治理、标准化建设作出刚性部署，为本文件编制提供清晰政策遵循。

《新一代人工智能发展规划》（2017，国务院）明确提出，“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”。这是国家层面首次将“伦理规范”“安全评估”纳入人工智能发展战略框架，强调了伦理风险管理的重要性。

《新一代人工智能伦理规范》（2021，科技部）提出“将伦理道德融入人工智能全生命周期”以及“促进公平公正、保护隐私安全、确保可控可信、强化责任担当”等六项基本伦理要求，为实现伦理治理提供了明确指导。

《关于加强科技伦理治理的意见》（2022，中央办公厅、国务院办公厅）强调“制定……人工智能等重点领域的科技伦理规范、指南等，完善科技伦理相关标准”“推动将重要的科技伦理规范上升为国家法律法规”，凸显了智能机器人伦理标准立项的法理依据和政策需求。

《国家人工智能产业综合标准化体系建设指南》（2024年，工业和信息化部等）强调“规范人工智能全生命周期的伦理治理要求，包括人工智能伦理风险评估，人工智能的公平性、可解释性等伦理治理技术要求与评测方法，人工智能伦理审查等标准”，为智能机器人伦理治理标准的制定与落地提供了具体技术指引和实践遵循。

《国务院关于深入实施“人工智能+”行动的意见》（2025，国务院）提出“深入研究人工智能对人类认知判断、伦理规范等方面的深层次影响和作用机理”“完善人工智能法律法规、伦理准则等”，为智能机器人伦理标准建设提供了顶层设计与实施路径。

《人工智能拟人化互动服务管理暂行办法》（2026，国家网信办、国家发展改革委、工信部、公安部、市场监管总局）提出“提供拟人化互动服务，应当遵守法律、行政法规，尊重社会公德和伦理道德”“拟人化互动服务提供者应当落实拟人化互动服务安全主体责任，建立健全算法机制机理审核、科技伦理审查、信息内容管理、网络和数据安全、风险预案和应急处置等管理制度”。

《人工智能科技伦理审查与服务办法（试行）》（2026，工信部、国家发展改革委、国家网信办等十部门）提出人工智能科技伦理治理是人工智能治理的重要组成部分，包括增进人类福祉、公平公正、可控可信、透明可解释、责任可追溯、隐私保护等内涵。加强人工智能科技伦理治理，是坚持科技向善、筑牢科技安全底线、保障产业高质量发展的必然要求。

（三）标准化背景：现行标准体系存在维度缺失、场景适配不足多重短板

国际层面，ISO、IEC、IEEE 等标准化组织及欧盟、美国、日本、韩国等主体虽发布通用人工智能伦理、风险管理标准与 AI 监管指引，但均面向通用线上 AI 系统，未结合智能机器人人机交互、多机协同的独有特性制定专项规范，缺少覆盖机器人全生命周期的完整伦理风险防控体系，难以适配国内机器人产业部署需求。国际通用 AI 伦理规范未适配机器人多机协同场景。

国内人工智能、机器人、风险管理领域已形成基础标准框架，但整体供给存在明显短板。GB/T 12643—2025《机器人 词汇》、GB/T 41867—2022《信息技术 人工智能 术语》仅界定基础术语，未明确人机交互、多机协同、伦理风险等细分概念；各类机器人安全国标侧重机械、网络硬件防护，对算法歧视、人机情感异化、群体涌现等伦理问题覆盖不足；GB/T 27921—2023《风险管理 风险评估技术》为通用风险管理标准，无法适配多机协同传导风险；通用 AI 伦理文件仅聚焦单一人机交互，缺失机器间协同、竞争带来的系统性风险管控内容，少量行业伦理规范也仅局限于医疗、金融少数领域。国内现有标准侧重硬件安全，缺乏

多机协同治理规则。

总而言之，目前国内外均无同时覆盖人机交互、多机协同双维度的智能机器人伦理风险管理专项标准，全球暂无统筹人机交互、多机协同双维度的专项风险管控标准。行业现有管控手段均以个体设备自查为主，缺少多机协同风险专用识别、处置工具，大规模协同场景易出现风险漏判、处置滞后。

（四）必要性和意义

夯实智能治理制度底座，填补标准空白。落实国家科技伦理治理部署，首创人机交互、多机协同双维度分层治理范式，将国家人工智能、科技伦理顶层政策要求转化为可执行、可操作的标准化操作流程。系统划分七类人机交互风险、五类多机协同风险，打通全域防控网络，补齐现有标准缺失多机协同治理规则的短板，完善“基础术语—硬件安全—网络安全—伦理风险管理”一体化机器人标准体系，助力我国人工智能现代化治理建设。统一行业智能机器人伦理风险整体管控框架，解决当前行业风险识别零散、评估依据缺失、防控措施各行其是的突出问题，响应国家规范具身智能向善发展的战略部署。

统一科技监管评判依据，提升治理效能。强化行业监管技术支撑，明晰多元主体权责，提升行业治理规范化水平。标准清晰界定智能机器人全生命周期风险管理核心，明确人机交互事故、多机协同级联损害两类场景下的风险溯源与责任判定流程，构建全链条权责溯源机制，形成客观量化的行业审查标尺，有效化解事故责任推诿、追责无据的现实难题；同时为主管部门开展伦理合规审查、风险事件处置、常态化监督抽查提供统一评判标尺，

为常态化伦理核查、重大风险处置提供标准化制度支撑，扭转以往监管标准模糊、监管实施无据可依的局面，全面提升智能机器人行业治理规范化、标准化水平。

驱动产业伦理合规升级，筑牢发展根基。破解产业双重治理痛点，分层精准管控风险，降低全产业链合规成本。针对人机交互应用场景，统一人机交互伦理风险识别、分析、评估、处置全流程规范，解决企业在算法公平、隐私保护、人机事故权责划分等方面缺少统一操作依据的痛点；面向多机协同场景，首创完整多机协同治理框架，规范任务分配、跨机数据互通、级联风险阻断、群体异常行为约束机制，填补多机协同系统性伦理管控空白。标准区分人机交互直接侵害风险与多机协同传导次生风险，对人机交互风险设置优先级，配套多机协同级联评估、故障熔断隔离机制，改变传统粗放式单一管控模式，消除大规模多机协同系统性隐患，推动伦理审查前置研发环节，兼顾人机交互、多机协同两类场景伦理风险管理需求，有效降低全产业链跨场景伦理合规成本，统筹技术创新与伦理底线，增强全产业链整体合规竞争力，推动技术创新与伦理约束协同发展，提升国内机器人产业核心竞争力，引领智能机器人向上向善发展。

全方位保障公众合法权益，平衡产业创新与公共利益，提升社会公众信任度。通过标准化手段约束算法歧视、隐私过度采集、机器隐性合谋、人机情感异化等各类伦理失范行为，重点保障老年人、儿童、残障人士等特殊群体平等使用智能机器人的合法权益，有效遏制公共空间挤占、大范围隐私泄露、差异化不公服务等社会负面问题，兼顾产业技术创新发展与民生公共福祉，切实

维护自然人生命安全、人格尊严与财产权益，提升群众对智能机器人产品的安全感、认同感。

构建全生命周期闭环治理机制，坚持伦理前置长效治理思维，引领产业高质量可持续发展。本文件覆盖智能机器人研发设计、生产制造、部署运维、回收处置全生命周期，建立“风险识别—分析—量化评估—分级控制—动态监测—持续改进”完整闭环管理体系，配套标准化风险自查清单、政务/养老/医疗/金融等高风险场景管控示例，引导市场主体将伦理审查前置至产品设计源头，从根源上削减伦理风险隐患，形成常态化、可持续的自律治理机制，为智能机器人产业长期安全、合规、高质量发展提供长效制度支撑。

综上所述，该标准的编制既响应相关政策导向，也是产业发展需求的必然选择，具有明确现实意义和战略价值。

二、工作简况

（一）前期准备

2025年10月，由深圳市福田区政务服务和数据管理局牵头联合相关代表性企业、科研院所等成立标准编制组，对本文件起草的必要性、可行性以及标准框架内容进行研讨。

（二）调研和草案编制阶段

2025年11月—12月，编制组系统调研国内外伦理治理、人工智能风险管理、机器人安全等相关政策、标准与文献，梳理产业实践痛点与监管需求，结合智能机器人全生命周期特点，多次召开内部工作组会议，确定标准框架与核心条款。编制组围绕风险管理原则、风险识别、分析、评估、控制、监测改进等内容细

化文本，形成标准初稿。

2026 年 4 月—5 月，编制并发放调研问卷，共回收有效问卷 9 份。调研内容涵盖行业背景、业务布局、伦理风险管理与治理实践、标准需求等多个维度，旨在为标准的编制提供实践依据和参考方向。

2026 年 5 月—6 月，编制组先后赴优必选、今日人才进行调研，并与深开鸿、信通院等企业、院所沟通，进一步完善标准内容。

（三）立项阶段

2026 年 4 月，编制组完成立项建议书并提交深圳市深圳标准促进会审核，深圳市深圳标准促进会于 2026 年 4 月批准本标准立项，计划完成时间为 2026 年 10 月。

三、标准主要内容的编制依据以及与国内领先、国际先进标准的对标情况

（一）编制依据

标准制定前编制组查阅和搜集了国内相关的标准、文献资料，调查了相关行业的概况，并在制定过程中多次与业内人士进行咨询和讨论，广泛征求意见。**标准化方面**，严格遵循 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》要求，规范标准体例、术语表述、章节逻辑、附录设置，保证文件格式合规统一。**术语与条款方面**，引用 GB/T 12643—2025《机器人 词汇》界定机器人基础定义；引用 GB/T 41867—2022《信息技术 人工智能 术语》统一人工智能相关概念；引用

GB/T 27921—2023《风险管理 风险评估技术》搭建通用风险管理流程；参考 GB/T 45081—2024《人工智能 管理体系》完善全生命周期管理要求。**政策参考方面**，深度融合《新一代人工智能伦理规范》《关于加强科技伦理治理的意见》《国家人工智能产业综合标准化体系建设指南》等国家政策文件的伦理核心要求，将以人为本、公平公正、隐私安全、可控可信、责任明确等顶层原则转化为标准可落地技术条款。**产业实践方面**，编制组调研养老、医疗、政务、金融、物流等多场景智能机器人企业、系统集成商、运营机构、监管单位，梳理人机交互、多机协同真实伦理风险案例，区分两类维度风险特征，分别制定适配的识别方法、评估模型、管控措施，保障标准实用性。

（二）与国内领先、国际先进标准的对标情况

目前国内外无专门针对智能机器人伦理风险管理的专项标准，现有相关文件分为三类：一是通用人工智能伦理规范，仅侧重人机交互，未覆盖多机协同风险；二是机器人安全标准，聚焦硬件、网络安全，缺少算法伦理、情感异化、多机协同级联风险管控内容；三是通用风险管理标准，未适配机器人实体交互、多智能体协同的独有场景特征，无法区分人机交互、多机协同两类差异化风险。

本文件为国内首份同时构建人机交互、多机协同双维度的智能机器人伦理风险管理团体标准，且首创分层双维度治理框架，明确人机交互为核心管控维度、多机协同为补充管控维度，分别定义识别范围、风险类型、分析路径、管控手段，填补多机协同伦理风险标准化治理空白；构建三维风险评估模型，兼顾人机交

互直接危害与多机协同级联传导危害，配套量化综合评价指数计算方法，实现风险分级量化；打通研发、生产、部署、运维、回收全生命周期，针对两类维度风险分别设计分级、分类管控措施，配套资料性风险清单、典型场景示例，可操作性强。本标准术语、基础管理流程与现行机器人、人工智能、风险管理国家标准完全衔接，无冲突条款；对于多机协同新增技术要求，作为现有标准的补充延伸，可与现有安全、数据、算法规范配套协同使用。

四、主要条款的说明以及主要技术指标、参数、试验验证的论述

本文件共包括 9 章正文、2 个资料性附录及参考文献，核心技术内容围绕人机交互、多机协同双维度分层设计，各章节编制思路与内容说明如下：

（一）范围

本文件给出了智能机器人伦理风险管理的原则和过程，提供了从人机交互和多机协同两个维度开展风险识别、分析、评估、控制、监测和改进等的具体内容。

本文件适用于智能机器人伦理风险管理活动。其中，人机交互维度适用于智能机器人设计开发、生产制造、部署应用、运行维护及回收处置等全生命周期的伦理风险管理；对于无物理实体的机器系统，其全生命周期包括设计开发、系统构建、部署上线、运行维护及系统退役与数据销毁等阶段，可参照本文件相应环节开展伦理风险管理。多机协同维度侧重于多机器人系统协同运行阶段的伦理风险管理，同时兼顾协同协议设计、系统架构规划、通信接口定义等设计开发环节的前置风险防控。

（二）规范性引用文件

本章列明标准编制过程中所引用的相关标准。

（三）术语和定义

本章除采用 GB/T 12643—2025、GB/T 41867—2022 中已界定的通用词汇外，在此基础上新增人机交互、多机协同、智能机器人伦理风险等适配双维度治理框架的专属术语并明确定义，统一行业对两类交互场景、伦理风险的认知口径，为标准落地实施提供统一语义基础。

（四）风险管理原则

本章确立以人为本、责任明确、公平公正、公开透明、隐私安全、可控可信六大核心风险管理原则，每条原则同步配套人机交互、多机协同双重约束要求，从顶层设计层面统筹人机交互、多机协同两类场景的伦理治理导向，为全流程风险管控提供价值遵循。

（五）风险识别

本章规定风险识别总体要求、识别范围与可选工具方法，创新划分人机交互、多机协同两大独立识别维度并设置差异化识别视角，分类梳理 7 类人机交互伦理风险与 5 类多机协同伦理风险，配套全生命周期风险分布表，对应附录 A 提供标准化风险清单模板。

（六）风险分析

本章分别针对人机交互、多机协同开展风险根源拆解，同步从发生可能性、危害严重程度、风险传导关联性三个维度开展系统性分析，为后续风险评估、分级管控提供完整基础研判依据。

（七）风险评估

本章构建“可能性—严重程度—关联性”三维度风险评估模型，配套智能机器人伦理风险综合评价指数计算公式；依据综合评价结果将伦理风险划分为四级管控等级，实现人机交互、多机协同两类风险统一量化、分级判定。

（八）风险控制

本章明确风险分级管控总体策略，按照人机交互、多机协同分设独立分类管控小节，针对各类风险分别给出技术、数据、制度、系统多层级管控措施；同步建立覆盖研发到回收的全生命周期闭环管控机制，对于机器系统（如数智员工、智能客服等）补充了 API 接口管控、模型权重销毁等针对性措施。

（九）风险监测和改进

本章规定人机交互、多机协同两类风险常态化动态监测机制，明确风险预警、重大事件应急处置、定期复盘评估要求；建立多方协同伦理审查机制，配套从业人员伦理能力建设、标准动态优化等持续改进措施，形成长效治理闭环。

（十）附录

附录 A 分别提供人机交互伦理风险清单和多机协同伦理风险清单，便于企业参考开展风险排查；附录 B 选取政务、养老、金融、医疗、仓储五类高频高风险应用场景，明确各场景专属伦理管控重点，为行业场景落地提供实操参考。

五、是否涉及专利等知识产权问题

无。

六、重大分歧意见的处理经过和依据

无。

七、实施标准的措施建议

本文件发布实施后，深圳市深圳标准促进会秘书处将及时向相关企业、科研机构、应用单位及监管部门通报标准发布信息，积极开展宣贯培训，推动各方理解、应用本文件。通过宣贯与实施，引导行业规范开展伦理风险管理，提升智能机器人安全可信水平，促进产业健康有序发展。

八、其他应予说明的事项

无。