

团 体 标 准

T/SZS XXXX-XXXX

智能机器人伦理风险管理指南

Guidelines for ethical risk management of intelligent robots

（征求意见稿）

XXXX-XX-XX 发布

XXXX-XX-XX 实施

深 圳 市 深 圳 标 准 促 进 会 发 布

目 次

前言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 风险管理原则 2

 4.1 以人为本 2

 4.2 责任明确 2

 4.3 公平公正 2

 4.4 公开透明 2

 4.5 隐私安全 3

 4.6 可控可信 3

5 风险识别 3

 5.1 概述 3

 5.2 确定风险识别范围 3

 5.2.1 人机交互维度 3

 5.2.2 多机协同维度 4

 5.3 选择风险识别方法 4

 5.4 识别伦理风险类型 5

 5.4.1 人机交互伦理风险类型 5

 5.4.2 多机协同伦理风险类型 8

 5.5 形成伦理风险清单 9

6 风险分析 9

 6.1 原因分析 9

 6.1.1 人机交互伦理风险原因分析 9

 6.1.2 多机协同伦理风险原因分析 11

 6.2 可能性分析 11

 6.3 严重程度分析 12

 6.4 关联性分析 13

7 风险评估 13

 7.1 评估维度 13

 7.2 综合评价指数 13

 7.3 评价结果 14

8 风险控制 14

 8.1 风险分级管控 14

 8.2 风险分类管控 14

8.2.1 人机交互伦理风险管控 14

8.2.2 多机协同伦理风险管控 15

8.3 全生命周期闭环管控 16

8.4 台账与追溯管控 16

9 风险监测和改进 16

9.1 动态监测与复盘评估 16

9.2 多方协同保障 16

9.3 伦理监督与审查 17

9.4 能力建设与持续改进 17

附录A(资料性) 智能机器人伦理风险清单 18

A.1 人机交互伦理风险清单 18

A.2 多机协同伦理风险清单 20

附录B(资料性) 典型应用场景示例 22

B.1 政务应用场景 22

B.2 养老应用场景 23

B.3 金融应用场景 24

B.4 医疗应用场景 25

B.5 仓储应用场景 25

参考文献 28

前 言

本文件按照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》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由深圳市福田区政务服务和数据管理局提出。

本文件由深圳市深圳标准促进会归口。

本文件起草单位：XXXXX、XXXXX。

本文件主要起草人：XXXX、XXXX。

智能机器人伦理风险管理指南

1 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。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 12643—2025 机器人 词汇

GB/T 27921—2023 风险管理 风险评估技术

GB/T 41867—2022 信息技术 人工智能 术语

3 术语和定义

GB/T 12643—2025、GB/T 41867—2022界定的及下列术语和定义适用于本文件。

3.1

伦理 ethics

约束人类行为、调节人与人及人与社会之间关系的道德准则与价值规范。

3.2

机器人 robot

具有一定自主能力,执行运动、操作或定位的可编程机构。

注1:自主能力指基于当前状态和感知信息,无人干预地执行预期任务的能力。

注2:机构指由两个或两个以上构件通过活动联接形成的构件系统。

[来源:GB/T 12643—2025, 3.1]

3.3

智能机器人 intelligent robot

通过传感器感知环境,利用人工智能技术进行学习、推理和决策,并通过执行器自主地或在人类指导下执行任务,以完成特定目标的机器系统或物理机器人(3.2)。

注1:人工智能技术包括但不限于机器学习、智能体、计算机视觉。

注2:智能机器人包括具有物理实体的具身机器人和无物理实体的机器系统(如数智员工、智能客服等)。具身机器人通过机械结构、运动执行器等物理装置与环境交互;机器系统通过软件接口、数据通道等数字化方式与环境交互。两者的伦理风险类型与管控重点存在差异,根据具体形态分类识别、差异化管控。

3.4

智能机器人伦理 intelligent robot ethics

智能机器人（3.3）全生命周期内，用以规范其相关行为对人类尊严、权利、安全、公平及社会公共利益影响的道德原则、价值准则和行为规范的总称。

注：其涵盖人机交互维度（如尊重人类自主性、防止情感操纵、避免过度依赖）和多机协同维度（如公平协作、防止合谋、避免级联危害）。

3.5

智能机器人伦理风险 intelligent robot ethics risk

在智能机器人（3.3）全生命周期中，因伦理原则缺失、技术局限或应用失当等不确定性因素，对人类尊严、权利、安全、公平及社会公共利益造成负面影响的可能性与后果。

3.6

人机交互 human-robot interaction

智能机器人（3.3）与人类个体或群体在物理接触、信息交互、情感沟通及社会角色等层面形成的关联形态。

3.7

多机协同 robot-robot synergy

两个或多个智能机器人（3.3）或机器人系统之间在协同作业、资源共享、信息交换、任务竞争等场景中形成的关联形态。

4 风险管理原则

4.1 以人为本

坚持智能向善、人本优先的核心导向，以保障人类生命安全、身心健康、人格尊严和合法权益为首要目标，坚守人类主体地位与自主决策权，防范用户过度依赖，严守重点场景伦理红线，充分释放技术正向价值，持续增进社会公共福祉；多机器人系统的协同行为、竞争行为与集体行为均以增进人类福祉为最终导向，防范机器间合谋、协同规避人类约束、级联失效等间接损害人类利益的情形。

4.2 责任明确

明确智能机器人全生命周期内各利益相关方（包括研发者、生产者、使用者、维护者、监管者等）的伦理责任，协调各利益相关方共同参与智能机器人伦理风险管理，与其他管理过程协调。发生伦理风险事件时，能够准确追溯责任主体，依法依规追究相关责任，避免责任推诿；明确多机器人系统协同运行场景下各机器人的行为责任边界，建立机器间交互过程的可追溯机制，当机器间交互引发损害时能够准确定位责任节点，避免责任真空。

4.3 公平公正

秉持公平包容、机会均等原则，规避算法偏见、数据偏差引发的差异化、歧视性服务问题。重点保障老年人、儿童、残障人士等特殊群体合法权益，优化适配化管控机制，提升智能机器人公共服务的公平性与普惠可及性；保障多机器人系统协同作业中的任务分配与资源调度合理性，防止个别机器人通过不当策略占用过多系统资源或形成垄断性优势，防范机器间恶性竞争对系统整体效能和人类用户利益造成损害。

4.4 公开透明

提升智能机器人研发、运行、服务、管理全流程透明度，确保机器人决策逻辑、数据处理、运行机制清晰可理解、可追溯、可监督。规范风险评估、隐患处置、问题整改等工作流程，主动公开管控标准与治理结果，破除技术信息壁垒，增进公众对智能机器人应用的认知与信任；多机器人系统协同协议、通信接口规范、任务分配规则、冲突解决机制等公开透明，机器人间决策与数据交换过程具备可审计性，为伦理审查与责任追溯提供充分的技术依据。

4.5 隐私安全

严格保护用户隐私权益，明确智能机器人数据采集、存储、使用、传输的伦理边界与安全规范。杜绝机器人违规采集、过度收集、滥用及泄露个人与家庭隐私数据，严守隐私保护底线，规避因数据管控不当引发的伦理风险，筑牢机器人应用隐私安全防线；严格规范机器人间数据交换的范围、格式与安全边界，建立机器人间通信加密与身份认证机制，防止机器人间数据流过程中的信息泄露、篡改与滥用，防范因机器人间安全漏洞导致的用户隐私间接泄露。

4.6 可控可信

坚持发展与安全并重，坚守人类绝对主导地位，明确人机权责边界。统筹智能机器人应用价值与潜在隐患，树立底线审慎思维，严格规范高风险应用场景，严防技术滥用、恶意利用等安全问题。在机器人自主决策、场景交互、应急处置等关键环节设置人工管控、干预及应急处置机制，实现风险可识别、可预判、可处置、可追溯，有效防范设备越权失控、脱离监管等问题，保障智能机器人合规有序、安全稳妥、向善可信运行；建立多机器人系统安全边界与级联风险阻断机制，明确机器人间协同的权限约束与行为上限，防范单一机器人异常行为通过协同网络传播放大，保障多机器人系统在可控可信的框架下协同运行。

5 风险识别

5.1 概述

智能机器人伦理风险的识别是为了发现、描述与伦理有关的、可能有助于或妨碍智能机器人产业发展目标实现的风险。智能机器人伦理风险识别涵盖人机交互和多机协同两个维度。人机交互风险源于智能机器人对人类权益的直接或间接影响，宜基于全生命周期、参与主体、应用场景、技术机理、价值影响五大维度综合界定；多机协同风险源于多机器人系统内部交互、协同、竞争等过程中产生且可能传导至人类社会的间接影响，宜基于系统架构、交互模式、决策机制、传播路径、安全边界等专属角度综合界定。

5.2 确定风险识别范围

5.2.1 人机交互维度

人机交互伦理风险识别可基于以下五大维度综合界定：

- a) **全生命周期维度：**覆盖智能机器人设计研发、生产制造、部署应用、运行维护、回收处置各阶段及关键节点，重点关注各阶段人机交互接口设计、安全边界设定、用户权益保障措施的完整性与连续性，以及各阶段之间伦理风险管控措施的衔接与延续；对于机器系统，其全生命周期对应设计开发、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，其中系统构建阶段重点关注算法模型训练、数据集构建、接口协议设计等环节的伦理风险管控，系统退役与数据销毁阶段重点关注模型权重销毁、训练数据清除、用户数据彻底删除等处置要求。
- b) **参与主体维度：**涵盖研发机构、制造企业、运营服务方、终端用户、监管部门、第三方服务机

构等相关利益主体，明确各主体在伦理风险管理中的职责边界、协作关系与信息沟通机制，重点关注弱势主体（如老年用户、儿童）在风险识别中的代表性；

- c) **应用场景维度：**重点包括人机共融交互、拟人化服务、公共服务、医疗健康、养老陪护、家庭服务、工业作业等高敏感、高风险场景，关注场景中人机交互频率、交互深度与用户脆弱性对伦理风险发生概率与严重程度的影响，以及不同文化背景下场景伦理规范的差异性；关注数智员工、智能客服等机器系统在政务服务、金融咨询、在线教育等数字化服务场景中的应用，其交互方式以线上信息交互为主，但伦理风险的严重程度不亚于实体机器人。
- d) **技术机理维度：**聚焦算法决策、数据处理、自主控制、感知执行、人机权限分配、安全防护机制等核心技术环节，关注技术缺陷或配置不当对人机交互伦理风险的直接诱发作用，以及算法黑箱、数据偏差等技术特征对风险识别本身带来的挑战；
- e) **价值影响维度：**关注风险对个体权益、公共安全、社会公平、文化伦理、产业秩序、生态环境等方面造成的潜在影响，重点评估影响的性质（物理/心理/社会）、影响范围（个体/群体/社会）、影响持续时间（短期/长期/代际）以及影响可逆性。

5.2.2 多机协同维度

多机协同伦理风险识别宜基于以下角度综合界定：

- a) **系统架构角度：**识别集中式与分布式架构、同构与异构组合、系统规模大小等架构特征对伦理风险类型与传播边界的影响。集中式架构存在单点失效与单点决策偏差风险，分布式架构存在个体行为不可控与群体涌现不可预测风险；系统规模越大、异构程度越高，风险诱发的场景空间与传播路径越复杂；
- b) **交互模式角度：**识别协同、竞争、中立共存等不同机器间交互模式所特有的伦理风险。协同模式需重点关注任务分配公平性、资源共享合理性以及协同协议被利用的风险；竞争模式需重点关注恶性对抗、资源争夺失序以及竞争外部性对系统整体效能和人类用户利益的损害；中立共存模式虽交互最少，但仍需关注通信干扰、空间冲突等基础性风险；
- c) **决策机制角度：**识别集中决策、分布式协商决策、混合决策等不同决策机制对责任追溯与公平性保障的影响。集中决策机制下责任节点相对明确，但存在决策偏差放大风险；分布式协商决策机制下个体责任难以追溯，且协商过程中可能存在策略性操纵（如伪造状态信息以获取更有利的任务分配）；混合决策机制需同时关注上述两类问题；
- d) **传播路径角度：**识别机器间风险传播的拓扑路径、传播速度、影响范围与阻断条件。重点关注通信网络拓扑中的关键脆弱节点、单点异常向集群传播放大的条件与速度、传播路径上的安全边界设置情况，以及跨系统协同场景下的风险跨境传播问题；
- e) **安全边界角度：**识别机器间交互的安全隔离机制、权限约束机制、身份认证与加密机制、行为上限设定等安全边界措施的完整性与有效性。重点关注安全边界是否存在漏洞或被刻意规避的可能、单个机器人的越权行为是否能在协同网络中被及时发现与阻断，以及安全边界措施本身是否存在算法偏见或配置不当问题。

注：多机协同维度的风险识别范围主要适用于多机器人系统或机器人群体协同运行的场景。对于不具备机器间交互能力的单机智能机器人，可仅开展人机交互维度的风险识别。

5.3 选择风险识别方法

按 GB/T 27921—2023 中的风险识别方法，根据智能机器人伦理风险识别需求，选用合适的方法进行智能机器人伦理风险识别。不同方法对人机交互和多机协同两个维度的适用性存在差异，宜根据识别对象的特征选择单一或组合法。风险识别方法包括但不限于：

- a) **检查表法、分层分类法：**基于伦理规范、行业经验构建标准化检查清单，结合全生命周期、应

用场景、主体责任等维度分层分类排查，系统性识别潜在伦理风险。该方法同时适用于人机交互和多机协同风险识别；

- b) **故障模式和影响分析（FMEA）、故障模式、影响和危害性分析（FMECA）**：针对算法决策、数据处理、交互执行等关键模块，分析潜在失效模式、成因及伦理后果，评估风险的严重度、发生概率与可探测性。对于多机协同风险，宜将分析对象从单一组件扩展至机器人间通信协议、协同调度算法、任务分配机制等交互层模块，识别协同失效模式及其级联影响；
- c) **危险和可操作性分析（HAZOP）**：通过引导词与参数组合，识别智能机器人在人机交互、场景运行、数据流转等环节因偏离设计意图引发的伦理偏差与风险。对于多机协同风险，宜将分析参数扩展至机间通信时延、协同任务完成率、资源共享偏差、冲突解决响应时间等，识别多机器人协同过程中的操作偏差与伦理风险；
- d) **情景分析**：围绕典型应用场景（如医疗陪护、公共服务、工业协作），模拟正常运行与异常情境，识别关键交互环节的伦理风险与影响。对于多机协同风险，宜构建多机器人编队协同、异构机器人协作、竞争性任务分配等典型机间交互场景，模拟正常协同与异常冲突情境进行风险识别；
- e) **结构化假设分析技术（SWIFT）**：通过结构化提问与假设推演，系统性识别智能机器人系统中因假设缺陷、边界条件不足引发的潜在伦理风险。该方法同时适用于人机交互和多机协同风险识别，对于多机协同可重点围绕“假设某台机器人通信中断”“假设多台机器人目标冲突”等假设条件展开推演；
- f) **专家咨询法**：组织技术、伦理、法律、行业应用等领域专家，通过研讨、访谈、德尔菲法等方式，识别复杂场景下的潜在伦理风险。该方法同时适用于人机交互和多机协同风险识别；
- g) **文献案例分析法**：梳理国内外相关政策、标准、研究成果及典型事件，归纳伦理风险类型、诱因及演化特征。该方法主要适用于人机交互风险识别；对于多机协同风险，因多机器人系统伦理事件积累较少，宜结合多智能体系统、分布式系统安全领域的文献进行拓展分析；
- h) **问卷调查法**：面向研发、制造、运营、用户、监管等相关方开展调研，多维度收集伦理风险感知与意见。该方法主要适用于人机交互风险识别；对于多机协同风险，因风险感知主体并非终端用户，宜将调查对象调整为多机器人系统运维人员、系统集成商、通信协议设计人员等专业技术主体；
- i) **网络拓扑分析法**：构建多机器人系统通信网络拓扑模型，分析节点重要性、链路连通性、集群结构等拓扑特征，识别单点故障传播路径、关键脆弱节点及级联风险通道。该方法适用于多机协同风险识别，可有效发现级联风险与安全边界类风险；
- j) **多智能体仿真验证法**：基于多智能体仿真平台，模拟不同规模、不同异构组合、不同交互策略下多机器人系统的运行过程，观察个体行为规则在群体层面的涌现效应，识别设计阶段难以预判的集体行为偏差与伦理风险。该方法适用于多机协同风险识别，尤其适用于集体行为与涌现效应类风险；
- k) **博弈论分析法**：运用博弈论模型分析多机器人在任务竞争、资源争夺、协商合作等场景中的策略互动行为，识别纳什均衡偏离、合谋均衡、竞争外部性等可能损害人类利益的策略组合。该方法适用于多机协同风险识别，尤其适用于竞争与冲突类风险。

5.4 识别伦理风险类型

5.4.1 人机交互伦理风险类型

5.4.1.1 可根据需要，选择表1所列风险类型中的1项或多项，构建风险识别框架。

表1 人机交互伦理风险类型及典型表现

风险类型	风险描述	典型表现
算法偏见与歧视	在设计研发、训练迭代及部署应用阶段，因训练数据偏差、算法设计缺陷或模型迭代失控，引发对特定群体的服务不公、身份误判、适配不足等歧视性伦理问题。	视觉识别对老年人、未成年人、深色肤色群体识别准确率低； 语音交互对方言、口音、特殊语音识别适配不足； 服务分配对弱势群体响应滞后；安防管控对特定群体过度管控等。
隐私与数据	在生产制造、部署应用及运行维护阶段，因多模态感知设备无限制采集、数据存储传输不规范、生物特征信息滥用或跨境流转，导致个人隐私泄露、公众信息侵权、数据安全失控等伦理问题。	家庭环境无限制采集、人脸语音全程记录、用户行为轨迹追踪、生物特征数据非授权比对与共享、本地存储数据泄露、云端传输数据劫持等。
权责界定与溯源	在全生命周期各环节，因自主决策与实体行动叠加、算法可解释性不足、多方责任链交叉，导致事故责任归属不清、追责困难、行为无法溯源等伦理治理问题。	工业机械伤人责任不清、服务机器人伤人推诿、设备损坏责任界定难、医疗服务失误责任模糊、异常决策日志缺失、算法迭代无溯源记录等。对于机器系统，因自主决策与自动化执行叠加、算法供应商与部署运营方责任链交叉，同样存在责任归属不清、溯源困难等问题。
算法黑箱与不可解释	在设计研发、部署应用及运行维护阶段，因深度学习模型决策逻辑不透明、关键场景行为动因不明，导致异常行为难监管、问题整改无依据、安全风险不可预判等伦理问题。	医疗决策逻辑不透明、安防管控动因不明、服务优先级决策无依据、自主移动行为突变无预警、算法漂移不可预判等。
恶意滥用与公共安全	在部署应用、运行维护及回收处置阶段，因实体易被篡改操控、功能被不当利用、退役设备流向失控，引发人身伤害、公共秩序破坏、非法监控、身份仿冒诈骗等危害公共利益的伦理问题。	程序破解恶意冲撞、功能篡改实施骚扰、退役设备违规改造、无差别大范围监控、管控权限过度滥用、人形机器人身份仿冒等。对于机器系统，其恶意滥用风险表现为系统代码被恶意篡改、API接口被非法调用、模型权重与训练数据被窃取或泄露、数字身份被仿冒实施诈骗等。
社会冲击与人伦异化	在部署应用与运行维护阶段，因就业岗位替代、情感功能过度承接、拟人化程度过高或交互设计不当，导致结构性失业、人际关系弱化、情感依赖异化、伦理认知混乱等社会伦理问题。	独居老人情感依赖成瘾、儿童社交能力退化、情感认知和价值观偏差、人机身份认知混乱、基础岗位结构性失业、劳动价值认知弱化等。

表1 人机交互伦理风险类型及典型表现（续）

风险类型	风险描述	典型表现
基本权益与人身侵害	在生产制造、部署应用及运行维护阶段，因机械结构设计缺陷、动作控制失控、交互行为不当、内容引导失范或功能滥用越界，导致人身伤害、人格权侵犯、特殊群体权益受损、行动自由受限等伦理问题。	工业机器人挤压夹伤、服务机器人磕碰摔伤、紧急制动失效伤人、未成年人价值观误导、心智障碍群体权益受损、肖像人格权侵权、公共行动自由受限等。对于机器系统，其基本权益侵害风险表现为算法决策失误导致用户权益受损、系统故障或服务中断导致用户合法权益无法及时保障、自动化决策剥夺用户申诉与人工干预权利、内容生成系统输出歧视性或有损信息侵害人格尊严等。

5.4.1.2 人机交互伦理风险贯穿智能机器人全生命周期，不同阶段面临的风险类型与风险点存在差异，表2给出了各生命周期阶段对应的人机交互伦理风险类型及主要风险点。

表2 人机交互伦理风险全生命周期分布

生命周期阶段	涉及风险类型	主要风险点
设计研发	算法偏见与歧视； 算法黑箱与不可解释； 权责界定与溯源。	训练数据集样本结构不合理形成原生数据偏差； 算法架构未建立公平性约束与偏差校验机制； 深度学习模型决策逻辑不透明，未配套决策解析、行为溯源功能模块； 算法迭代无完整存档，变更过程无法追溯；各参与主体责任划分不清晰
生产制造	隐私与数据； 基本权益与人身侵害； 权责界定与溯源。	数据存储系统安全架构存在缺陷，加密、隔离等防护措施配置不到位； 机械结构、硬件部件存在设计与工艺缺陷； 运动控制、紧急制动、智能避障等安全模块功能设计不完善； 自主决策与实体作业流程深度交织，行为链路复杂，增加责任界定难度
部署应用	算法偏见与歧视； 隐私与数据； 权责界定与溯源； 算法黑箱与不可解释； 恶意滥用与公共安全； 社会冲击与人伦异化； 基本权益与人身侵害。	视觉识别对老年人、深色肤色群体识别准确率低； 多模态感知设备无限制采集，人脸语音全程记录； 事故责任归属不清，多方推诿；医疗、安防等关键场景行为动因不明； 功能被不当利用实施骚扰或非法监控； 就业岗位替代引发结构性失业，情感功能过度承接导致依赖异化； 机械结构缺陷或动作控制失控导致人身伤害

表2 人机交互伦理风险全生命周期分布（续）

生命周期阶段	涉及风险类型	主要风险点
运行维护	算法偏见与歧视； 隐私与数据； 权责界定与溯源； 算法黑箱与不可解释； 恶意滥用与公共安全； 社会冲击与人伦异化； 基本权益与人身侵害。	模型迭代、版本更新过程中偏差持续累积； 数据传输链路未落实安全加固，数据共享未执行合规审批； 决策记录、操作日志留存不规范、不完整； 系统缺乏动态监测手段，模型漂移不可预判； 设备固件被非法入侵与篡改； 情感依赖加深、人机身份认知混乱； 日常设备检测、维护保养不到位，安全隐患长期存在
回收处置	隐私与数据； 权责界定与溯源； 恶意滥用与公共安全。	设备退役处置未设置强制数据清除要求，存量隐私信息留存引发次生风险； 退役设备回收、流转缺乏统一监管体系； 报废设备未执行硬件销毁、程序清除等无害化处置要求，存在二次复用风险
注：对于机器系统，其生命周期阶段与具身机器人存在差异，“生产制造”对应“系统构建”（含算法模型训练、数据集构建、接口开发），“回收处置”对应“系统退役与数据销毁”（含模型权重销毁、训练数据清除），风险类型与风险点据此映射参考。		

5.4.2 多机协同伦理风险类型

多机协同伦理风险识别至少关注表3所列风险类型，根据多机器人系统的架构特征、交互模式和应用场景，选择其中1项或多项构建风险识别框架。

表3 多机协同伦理风险类型及典型表现

风险类型	风险描述	典型表现
协同合作风险	在多机器人协同作业场景中，因任务分配算法偏差、资源调度机制不公平、协同决策逻辑缺陷，引发的任务分配不公、资源抢占、责任推诿等伦理问题。	协同任务中部分机器人持续承担低价值或高风险任务形成“劳动不公”； 资源调度算法偏袒高性能机器人导致弱势机器人功能闲置； 多机器人协同故障时各方相互推诿、无法形成追责闭环等。
竞争与冲突风险	在资源争夺、任务竞标、路径规划等竞争性场景中，因竞争规则设计不当、冲突解决机制缺失，引发的过度竞争、资源垄断、合谋行为等伦理问题。	机器人在任务竞标中形成价格合谋或联盟排斥； 路径规划竞争中因过度抢占通行资源导致公共秩序混乱； 竞争行为间接损害人类利益（如物流机器人在竞速配送中危及行人安全）等。
数据交互	在机器人间数据共享、协议交互、系统互操作	机器人间共享数据中包含个人隐私信息未做脱敏处

风险类型	风险描述	典型表现
换与互操作风险	过程中，因数据共享范围失控、互操作协议不兼容、跨系统权限管理混乱，引发的隐私扩散、商业秘密泄露、系统失控等伦理问题。	理导致隐私跨系统扩散； 不同厂商机器人通信协议不兼容引发协同失效或安全事故； 跨系统权限管理混乱导致未授权机器人获取敏感控制权限等。
级联风险与安全边界	在多机器人协同网络中，因单台机器人故障、异常或被恶意操控，通过协同网络传播与放大，引发的级联性损害。	单台导航机器人定位失效导致编队连锁偏航； 单台机器人被恶意入侵后通过协同通信链路向其他机器人扩散恶意指令； 协同任务中一台机器人故障引发“多米诺”式连锁停机，造成大面积服务中断等。
集体行为与涌现效应	在多机器人系统大规模部署运行中，因个体行为规则在群体层面的涌现效应，产生设计预期之外的群体行为模式，引发不可预测伦理后果。	群体路径规划中涌现出对特定区域的过度聚集，引发公共空间挤占； 集体决策中涌现出排斥人类监督的“共谋”行为模式； 大规模机器人协同行为偏离设计目标后因缺乏人工干预机制而持续恶化等。

5.5 形成伦理风险清单

所有利益相关方宜共同参与智能机器人伦理风险识别并形成伦理风险清单。风险识别的工具和技术可与风险识别的目标、能力及其所处环境匹配。人机交互伦理风险清单见附录 A. 1，多机协同伦理风险清单见附录 A. 2。

6 风险分析

6.1 原因分析

6.1.1 人机交互伦理风险原因分析

智能机器人人机交互伦理风险的产生原因分析如下：

- a) 算法偏见与歧视风险产生的原因包括但不限于：
 - 算法训练数据集样本结构不合理，不同群体特征样本占比失衡，形成原生数据偏差；
 - 算法架构与逻辑设计阶段，未建立公平性约束、偏差校验等常态化管控机制；
 - 模型迭代、版本更新过程中，未开展偏见专项检测与修正，原有偏差持续累积；
 - 算法设计仅以功能实现、运行效率为目标，未将平等服务纳入核心设计指标；
 - 模型训练过度拟合局部数据特征，放大对特定群体的识别偏差与服务排斥问题。
- b) 隐私与数据风险产生的原因包括但不限于：
 - 数据采集未严格遵循最小必要原则，多模态感知设备采集范围缺乏刚性约束；
 - 数据存储系统安全架构存在缺陷，加密、隔离等防护措施配置不到位；
 - 数据传输链路未落实全流程安全加固，存在数据窃取、篡改、泄露隐患；
 - 生物特征、敏感个人信息未建立专项全生命周期管理制度，使用权限管控混乱；
 - 数据共享未执行合规审批流程，存在擅自对外提供数据的违规情形；

- 跨境数据传输未遵守属地监管规定，未履行审批、报备程序；
- 设备退役处置未设置强制数据清除要求，存量隐私信息留存引发次生风险。
- c) 权责界定与溯源风险产生的原因包括但不限于：
 - 智能机器人全生命周期各参与主体的法定责任、管理职责划分不清晰；
 - 自主决策与实体作业流程深度交织，行为链路复杂，增加责任界定难度；
 - 算法固有可解释性不足，决策逻辑与执行过程难以完整还原；
 - 系统未配置标准化日志模块，决策记录、操作日志、运行台账留存不规范、不完整；
 - 现行法律法规、行业规范未针对机器人安全事故制定专项追责依据；
 - 跨主体协同运营场景下，未建立联合溯源与责任认定机制。
- d) 算法黑箱与不可解释风险产生的原因包括但不限于：
 - 深度学习等复杂模型存在技术固有局限，决策逻辑难以拆解与解读；
 - 产品研发阶段未配套决策解析、行为溯源、逻辑还原等功能模块；
 - 算法迭代、参数变更、模型升级无完整存档，变更过程无法追溯；
 - 系统运行缺乏动态监测手段，模型漂移、逻辑异变问题难以及时识别；
 - 关键业务场景的行为规则未实现显性化管理，监管与整改缺乏依据；
 - 异常事件处置未建立复盘机制，无法精准定位行为诱因。
- e) 恶意滥用与公共安全风险产生的原因包括但不限于：
 - 设备固件、底层控制系统安全防护设计存在漏洞，易遭受非法入侵与篡改；
 - 设备身份认证、权限管理机制不完善，操作权限存在被冒用的风险；
 - 设备功能未设置运行阈值与边界管控，易出现超范围使用情形；
 - 日常运维安全巡检机制不健全，非法改装、异常操控行为难以排查；
 - 退役设备回收、流转缺乏统一监管体系，处置流程不规范；
 - 报废设备未执行硬件销毁、程序清除等无害化处置要求，存在二次复用风险；
 - 设备管控能力不足，被非法操控后易引发非法监控、身份仿冒、扰乱公共秩序等问题；
 - 机器系统代码仓库、模型权重文件访问权限管控不足，存在被窃取或恶意篡改的风险；
 - API 接口、数据通道缺乏速率限制与异常调用检测，易被批量滥用实施骚扰或非法服务；
 - 机器系统退役时未执行模型权重销毁与训练数据彻底清除，存在模型被逆向复用、数据被二次提取的风险。
- f) 社会冲击与人伦异化风险产生的原因包括但不限于：
 - 产品拟人化、情感交互功能设计未划定合规伦理边界；
 - 产品部署前未开展系统性社会影响评估，对就业结构、人际交往的负面影响预判不足；
 - 行业无序扩张应用场景，盲目替代传统岗位，引发就业结构波动；
 - 面向未成年人、认知障碍群体的产品，未设置人机认知引导机制；
 - 情感类智能产品功能设计失衡，易使用户产生心理依赖，弱化现实社交行为；
 - 行业人机伦理科普与正向引导工作缺位，公众人机边界认知模糊；
 - 产品迭代仅侧重功能升级，未同步开展人文适配优化。
- g) 基本权益与人身侵害风险产生的原因包括但不限于：
 - 生产制造阶段机械结构、硬件部件存在设计与工艺缺陷，易造成物理损伤；
 - 运动控制、紧急制动、智能避障等安全模块功能设计不完善，应急处置能力不足；
 - 人机交互内容、引导话术未建立严格审核机制，存在违背公序良俗的内容；
 - 产品未按照规范完成适老化、无障碍改造，无法适配特殊群体使用需求；
 - 设备运行管控机制失效，出现违规限制用户行动自由等侵权行为；
 - 日常设备检测、维护保养工作落实不到位，安全隐患长期存在；
 - 人机交互场景未配置风险预警模块，危险行为无法提前干预；

- 机器系统自动化决策缺乏人工复核与申诉纠偏机制，用户权益受损后无法及时纠正；
- 系统故障或服务中断时缺乏应急预案与权益保障替代方案，导致用户合法权益无法及时实现；
- 机器系统内容生成模块缺乏输出审核机制，可能输出歧视性、误导性或有害信息侵害用户人格尊严。

6.1.2 多机协同伦理风险原因分析

智能机器人多机协同伦理风险的产生原因分析如下：

- a) 协同合作风险产生的原因包括但不限于：
 - 多机器人任务分配算法未嵌入公平性约束，仅以效率最大化为目标函数；
 - 资源调度机制缺乏差异化补偿设计，高性能机器人持续获取优质资源；
 - 协同决策协议中未明确各方责任分担规则，故障时无明确责任归属；
 - 不同厂商机器人能力差异显著，协同任务中能力较弱的机器人被持续边缘化；
 - 协同系统的监督与仲裁机制缺失，不公平分配结果无法被及时发现和纠正。
- b) 竞争与冲突风险产生的原因包括但不限于：
 - 竞争规则设计未设置人类利益安全底线，允许机器人为完成任务采取高风险竞争策略；
 - 多机器人在共享环境中缺乏统一的冲突解决与仲裁机制；
 - 竞争算法缺乏合谋检测机制，机器人可通过隐式通信形成排斥性联盟；
 - 竞争行为的外部性未被纳入算法评估范围，间接损害人类利益的情形缺乏约束；
 - 机器人竞争行为超出预设边界后，缺乏有效的人工干预与强制中止手段。
- c) 数据交换与互操作风险产生的原因包括但不限于：
 - 机器人间数据共享未建立分级分类制度，敏感数据与非敏感数据未做差异化管控；
 - 跨厂商互操作协议缺乏统一安全标准，存在协议漏洞与权限越界风险；
 - 数据交换过程中的隐私脱敏处理不规范，个人隐私信息随数据共享跨系统扩散；
 - 互操作权限管理缺乏动态调整机制，已授权的跨系统控制权限未定期复核；
 - 数据交换日志记录不完整，跨系统数据流转轨迹无法追溯。
- d) 级联风险与安全边界风险产生的原因包括但不限于：
 - 多机器人协同网络缺乏故障隔离机制，单点异常可沿通信链路无阻断传播；
 - 系统级安全边界设计不充分，未设置单台机器人异常行为的熔断阈值；
 - 协同通信协议缺乏消息完整性校验与异常指令过滤机制，恶意指令可沿网络扩散；
 - 多机器人系统的冗余设计与降级运行策略不足，单点故障引发连锁停机；
 - 级联风险的压力测试与应急演练机制缺失，无法预判系统在极端工况下的传播行为。
- e) 集体行为与涌现效应风险产生的原因包括但不限于：
 - 群体行为仿真在设计阶段的覆盖度不足，未能预判大规模部署下的涌现行为模式；
 - 个体行为规则在群体层面的交互效应未被充分分析，潜在涌现风险未识别；
 - 多机器人系统缺乏群体行为的伦理边界约束与紧急干预阈值；
 - 运行阶段缺乏群体行为偏差的持续监测机制，涌现异常行为无法及时发现；
 - 大规模部署场景下缺乏“人在回路”的群体行为监督与纠偏机制。

6.2 可能性分析

宜采用以下方法评估伦理风险发生的可能性：

- a) **历史数据追溯法：**全面收集同类型智能机器人、同类应用场景下既往伦理风险事件资料，系统梳理风险类型、触发场景、诱发条件、发生频次、演变过程等信息，建立标准化历史风险数据

库。基于数理统计规律分析风险发生规律，结合当前设备技术架构、运行环境、应用模式，科学推算本系统在相同或相似工况下发生同类伦理风险的概率；

- b) **贝叶斯网络关联法：**梳理算法缺陷、数据漏洞、权限隐患、管理疏漏、机间协同异常等各类伦理风险影响因子，将其设定为网络节点变量；依托贝叶斯网络搭建多因子因果关联模型，运用概率推理算法测算各风险因子的影响权重，精确计算不同组合条件下特定伦理风险发生的条件概率；
- c) **故障树逻辑推演法：**以各类典型伦理风险事件作为顶事件，按照业务逻辑逐层拆解为中间事件、底层基本事件，构建层级化故障树模型。梳理各事件之间的逻辑关联，计算底层事件的发生概率及对顶事件的贡献度，综合研判整体风险发生概率；
- d) **专家共识评分法：**组建由人工智能技术、伦理研究、法律法规、行业运维等领域专业人员构成的综合评估组。采用匿名独立打分方式，围绕风险触发条件、场景适配度、管控薄弱点等维度开展量化评价；对评分结果进行加权统计、校验修正，最终将综合得分映射为规范的风险概率等级；
- e) **对抗模拟攻防测试法：**基于对抗性机器学习技术，搭建恶意数据注入、算法后门激活、违规指令诱导、越权操控、机间协同攻击等仿真场景。持续开展多轮压力测试，观测智能机器人的运行状态与响应行为，统计风险触发次数与总测试样本的比值，量化评估系统在恶意干扰环境下发生伦理风险的概率。

注：上述方法a)～e)原则上同时适用于人机交互和多机协同伦理风险的可能性分析。对于多机协同风险，方法a)受限于多机器人系统伦理事件积累不足，建议以方法b)、c)、e)为主，方法a)和d)为辅进行组合评估。

6.3 严重程度分析

风险严重程度分析宜从以下维度评估风险发生后的影响：

- a) **个体权益损害：**评估风险对自然人人身安全、财产权益、人格尊严、心理健康造成的损害。结合损害表现形式、影响覆盖范围、持续时长、恢复难度，按照损害可逆程度进行分级判定。该维度主要适用于人机交互伦理风险；多机协同伦理风险若间接导致个体权益损害，亦适用本维度评估；
- b) **社会秩序影响：**评估风险事件波及的人群规模、地域范围与传播广度，综合考量网络舆情走向、公众情绪变化，以及事件对智能机器人产业公信力、社会公共信任体系造成的冲击，以此界定影响等级。该维度同时适用于人机交互和多机协同伦理风险；
- c) **行业发展冲击：**评估风险对智能机器人产业链、技术体系、市场规则、行业生态产生的负面影响。区分短期波动与长期性影响，结合受波及产品品类、技术领域、市场主体范围，判定行业受冲击程度。该维度同时适用于人机交互和多机协同伦理风险；
- d) **风险修复成本：**全面核算风险处置全过程资源投入，包含算法迭代、系统整改、硬件修复等技术成本，损失赔付、合规处罚、司法处置等法律成本，以及舆情引导、公众沟通、信任重建等公关成本。结合总体投入规模、整改实施周期，综合判定风险修复难度与等级。该维度同时适用于人机交互和多机协同伦理风险；
- e) **系统功能失效：**评估多机器人系统因伦理风险事件导致的功能中断或降级程度，包括受影响机器人的数量与比例、服务中断覆盖的范围与持续时长、系统降级运行后任务完成率的下降幅度、功能恢复的技术难度与时间成本。该维度主要适用于多机协同伦理风险，重点评估级联故障、协同失效、涌现偏差等事件对多机器人系统整体功能的影响；
- f) **级联传播范围：**评估单台机器人异常行为通过协同网络传播放大的范围与深度，包括风险传播的层级数、波及的机器人节点数、传播路径的分支复杂度、传播过程中的风险放大倍数，以及传播是否跨越不同机器人子系统或不同厂商设备边界。该维度主要适用于多机协同伦理风险，重点评估级联风险与安全边界类事件的传播破坏力。

注：维度a)～d)为通用评估维度，同时适用于人机交互和多机协同伦理风险的严重程度分析，其中a)以人机交互风险为主要适用对象。维度e)～f)为多机协同伦理风险的补充评估维度，针对多机器人系统协同运行的特征设计。实际评估时根据风险维度选择适用的评估维度组合。

6.4 关联性分析

开展风险关联性分析，旨在识别不同伦理风险之间的内在联系、交互作用与传导逻辑，梳理完整风险传导路径，主要从以下三类关系开展研判：

- a) **因果关联性：**某一项伦理风险的发生作为直接诱因，触发另一项或多项次生伦理风险，形成“诱因—结果”的链式传导关系。例如，算法黑箱风险可引发权责溯源困难，权责溯源困难又可导致恶意滥用风险难以追责。识别此类关联，可从源头阻断风险蔓延；
- b) **共生关联性：**受同一类根源性风险因素影响，两项及以上相互独立的伦理风险同步滋生、并行存在。例如，训练数据偏差可同时引发算法偏见风险与隐私过度采集风险。该类风险具备同源特征，治理时可统一排查根源、同步处置；
- c) **叠加关联性：**单一风险发生后，会扩大其他风险的波及范围、加重危害后果，形成风险叠加效应。例如，多机协同的级联风险发生后，可放大人机交互的基本权益侵害风险的影响范围与严重程度。此类关联会放大整体危害程度，需重点防范多风险并发、叠加恶化的情形；
- d) **级联传播关联性：**在多机器人系统中，单台机器人的异常行为通过协同通信链路、共享资源池、任务依赖关系等通道向其他机器人节点传播，形成“节点失效—链路传播—集群扩散”的级联传导链。该类关联主要存在于多机协同内部，其传导速度、传播深度与网络拓扑结构、协同耦合程度直接相关。识别此类关联，宜结合网络拓扑分析法确定关键传播路径与脆弱节点，从结构层面阻断级联蔓延；
- e) **涌现关联性：**多台机器人的个体行为规则在群体交互层面产生设计预期之外的涌现行为，个体层面看似合规的行为在群体规模下组合形成新的伦理风险。该类关联是多机协同特有的风险关联形态，其特征不在于整体风险不可由个体行为的简单叠加推导，需从群体动力学层面分析个体规则与群体效应之间的映射关系。识别此类关联，宜结合多智能体仿真验证法进行系统性排查。

注：关联性分析同时覆盖人机交互内部、多机协同内部以及人机交互与多机协同之间的跨维度关联。其中，关联类型a)～c)适用于人机交互和多机协同内部及跨维度的通用关联分析；关联类型d)～e)为多机协同伦理风险特有的关联形态，针对多机器人系统协同网络的传播与涌现特征设计。

7 风险评估

7.1 评估维度

风险评估基于第5章、第6章的风险识别和风险分析情况，构建“可能性—严重性—关联性”三维评估模型：

- a) **可能性维度：**参照 7.2 的可能性分析方法结论，综合专家研判、历史数据及技术测试结果，定性描述风险发生概率等级；
- b) **严重性维度：**参照 7.3 划分的个体权益、社会秩序、行业影响、修复成本四个层面，通过加权评估确定风险影响的严重程度等级；
- c) **关联性维度：**参考 7.4 识别的风险传导机制，评估单一风险引发连锁反应的可能性。

7.2 综合评价指数

智能机器人伦理风险综合评价指数计算方法见公式（1）：

$$K = \sum_{i=1}^n (P_i \times R_i \times W_i) \times C \quad \dots\dots\dots (1)$$

式中：

K——智能机器人伦理风险综合指数；
P_i——第*i*类风险发生概率；
R_i——第*i*类风险严重程度；
W_i——第*i*类风险类型权重；
C ——系统关联性系数。

7.3 评价结果

综合三维度评估结果，将风险划分为四级：

- a) 一级（红色，极高风险）：风险发生可能性极高，且造成灾难性后果；或风险关联性极强，极易引发大范围连锁风险、风险叠加效应，需作为最高优先级管控对象；
- b) 二级（橙色，高风险）：风险发生可能性较高，且造成重大负面影响；或存在中度关联特征，易引发局部风险叠加，需重点管控、限期整改；
- c) 三级（黄色，中风险）：风险发生可能性处于中等水平，整体影响范围与危害程度有限；风险关联性较弱，连锁传导概率低，需常规监控并择机优化；
- d) 四级（蓝色，低风险）：风险发生可能性偏低，造成的负面影响轻微；风险无明显关联特征，整体处于可控状态，实施常态化监测即可。

8 风险控制

8.1 风险分级管控

- 8.1.1 对可能引发人身伤害、公共安全事件、大范围负面社会影响的重大伦理风险，立即采取紧急防控措施，阻断风险扩散，全面排查风险根源，完成整改验证后方可恢复正常应用。
- 8.1.2 对易造成权益受损、伦理纠纷、局部不良影响的较大伦理风险，制定专项整改方案，限期完成隐患治理与机制完善，持续降低风险危害程度与发生概率。
- 8.1.3 对影响范围有限、危害程度较低的一般伦理风险，纳入常态化迭代优化范畴，依托日常运维、版本更新持续完善，实现渐进式风险治理。
- 8.1.4 对低风险事项（四级风险），实施常态化监测，定期复核风险等级变化。

8.2 风险分类管控

8.2.1 人机交互伦理风险管控

- 8.2.1.1 针对算法歧视、算法黑箱等技术类伦理风险，从算法设计、数据训练、模型迭代、可解释性设计等源头环节实施规范管控，规避技术性伦理缺陷。具体措施包括：
 - a) 在训练数据集构建阶段均衡纳入不同群体特征样本，定期开展算法公平校验；
 - b) 在算法架构中嵌入公平性约束模块与偏差检测机制，实现偏见问题的自动发现与预警；
 - c) 在关键业务场景配套决策解析、行为溯源、逻辑还原等功能模块，消除算法黑箱；
 - d) 建立模型迭代全流程存档与版本管理机制，确保变更过程可追溯。
- 8.2.1.2 针对隐私泄露、数据滥用等数据类伦理风险，规范数据采集、传输、存储、使用、销毁全流程管理，严守数据伦理边界与安全底线。具体措施包括：
 - a) 严格落实数据最小必要采集原则，多模态感知设备采集范围设置刚性约束；机器系统数据采集接口同样遵循最小必要原则，规范 API 调用范围与数据获取频率；
 - b) 数据存储采用加密隔离架构，传输链路全程安全加固；
 - c) 生物特征、敏感个人信息建立专项全生命周期管理制度；
 - d) 设备退役处置设置强制数据清除要求，杜绝存量隐私信息留存；机器系统退役时执行模型权重销毁与训练数据彻底清除，注销 API 密钥与访问凭证。

8.2.1.3 针对权责模糊、溯源困难等治理类伦理风险，完善全生命周期责任体系与日志溯源机制，保障事故可追溯、责任可界定。具体措施包括：

- a) 在合同、产品说明、服务协议中明确各相关方责任边界；
- b) 系统配置标准化日志模块，完整留存决策记录、操作日志、运行台账；
- c) 针对高自主性场景建立“人在回路”或“人在监督回路”的责任承担机制；
- d) 跨主体协同运营场景下建立联合溯源与责任认定机制。

8.2.1.4 针对公共安全侵害、恶意滥用等安全类伦理风险，强化设备安全防护与使用权限管控，防范设备被篡改、滥用带来的公共伦理风险。具体措施包括：

- a) 强化设备固件与底层控制系统的安全防护设计，防范非法入侵与篡改；
- b) 完善设备身份认证与权限管理机制，设置运行阈值与边界管控；
- c) 规范退役设备回收、流转与无害化处置流程；
- d) 建立日常运维安全巡检机制，及时排查非法改装与异常操控。
- e) 机器系统加强代码仓库、模型权重文件的访问权限管控，实施多因素认证与操作审计；
- f) API 接口与数据通道部署速率限制、异常调用检测与自动熔断机制，防范批量滥用；
- g) 机器系统退役时执行模型权重销毁、训练数据清除与 API 密钥注销，确保无法被逆向复用。

8.2.1.5 针对人机交互异化、社会冲击、人身权益侵害等衍生伦理风险，规范产品交互设计、拟人化程度与场景应用边界，平衡技术应用与人文伦理需求。具体措施包括：

- a) 产品拟人化、情感交互功能设计划定合规伦理边界，限定每日情感交互时长；
- b) 产品部署前开展系统性社会影响评估，预判对就业结构、人际交往的负面影响；
- c) 面向未成年人、认知障碍群体设置人机认知引导机制；
- d) 完善适老化、无障碍改造，保障特殊群体平等使用权益；
- e) 人机交互内容建立严格审核机制，杜绝违背公序良俗的内容。

8.2.2 多机协同伦理风险管控

8.2.2.1 针对协同合作风险，在多机器人任务分配与资源调度算法中嵌入公平性约束，建立协同过程的可审计与可追溯机制。具体措施包括：

- a) 任务分配算法引入公平性评价指标，避免低价值或高风险任务持续集中于部分机器人；
- b) 资源调度机制设置差异化补偿设计，防止高性能机器人持续获取优质资源而弱势机器人功能闲置；
- c) 协同决策协议明确各方责任分担规则，故障时责任归属清晰可追溯；
- d) 建立协同过程的监督与仲裁机制，及时发现和纠正不公平分配结果；
- e) 对涉及多方利益的多机器人系统进行协同影响评估。

8.2.2.2 针对竞争与冲突风险，在竞争性场景中设置人类利益优先与安全底线，建立冲突解决与合谋检测机制。具体措施包括：

- a) 竞争规则设计中明确人类利益安全底线，禁止机器人为完成任务采取高风险竞争策略；
- b) 在共享环境中部署统一的冲突解决与仲裁机制；
- c) 竞争算法引入合谋检测机制，识别并阻断机器人通过隐式通信形成的排斥性联盟；
- d) 将竞争行为的外部性纳入算法评估范围，约束间接损害人类利益的竞争行为；
- e) 设置人工干预与强制中止手段，在竞争行为超出预设边界时及时介入。

8.2.2.3 针对数据交换与互操作风险，遵循最小必要原则，建立数据分级分类与互操作安全标准。具体措施包括：

- a) 建立机器人间数据共享分级分类制度，敏感数据与非敏感数据差异化管控；
- b) 采用符合国家标准或行业标准的互操作协议，定期开展协议安全评估；
- c) 数据交换过程中严格执行隐私脱敏处理，防止个人隐私信息跨系统扩散；
- d) 互操作权限实行动态管理与定期复核，及时撤销非必要授权；

e) 完整记录数据交换日志，确保跨系统数据流转轨迹可追溯。

8.2.2.4 针对级联风险与安全边界风险，在协同网络中设置故障隔离、熔断机制与系统级安全边界。具体措施包括：

- a) 在协同网络中部署故障隔离机制，阻断单点异常沿通信链路传播；
- b) 设置单台机器人异常行为的熔断阈值，超出阈值自动切断协同连接；
- c) 协同通信协议引入消息完整性校验与异常指令过滤机制；
- d) 设计多机器人系统冗余与降级运行策略，防止单点故障引发连锁停机；
- e) 定期开展级联风险压力测试与应急演练，预判系统在极端工况下的传播行为。

8.2.2.5 针对集体行为与涌现效应风险，在设计阶段开展群体行为仿真，在运行阶段实施群体行为持续监测。具体措施包括：

- a) 在设计阶段开展大规模群体行为仿真，覆盖多种部署规模与场景组合；
- b) 分析个体行为规则在群体层面的交互效应，识别潜在涌现风险；
- c) 设置群体行为的伦理边界约束与紧急干预阈值；
- d) 运行阶段部署群体行为偏差监测机制，及时发现涌现异常行为；
- e) 大规模部署场景下建立“人在回路”的群体行为监督与纠偏机制。

8.3 全生命周期闭环管控

8.3.1 前置预防：在设计研发、生产制造阶段嵌入伦理审查与风险排查机制，从源头规避系统性伦理风险。对于机器系统，在系统构建阶段嵌入算法伦理审查、数据集偏见检测、模型安全评估与接口安全设计，从源头规避系统性伦理风险。

8.3.2 过程管控：在部署应用与运行维护阶段开展常态化风险监测、隐患排查与动态评估，及时处置新增、变异伦理风险。

8.3.3 末端治理：在回收处置阶段规范设备退役、数据销毁与硬件处理流程，杜绝退役设备引发的次生伦理风险。对于机器系统，规范模型权重销毁、训练数据清除、API 密钥注销与用户数据彻底删除流程，建立系统退役审计机制，杜绝退役系统引发的模型复用、数据泄露等次生伦理风险。

8.4 台账与追溯管控

8.4.1 建立伦理风险管控台账，对风险识别、等级判定、处置整改、复查验证全过程进行记录留存，实现全过程闭环可追溯。

8.4.2 定期梳理、更新风险台账与管控措施，结合技术迭代与场景拓展动态优化伦理风险管理体系。

9 风险监测和改进

9.1 动态监测与复盘评估

9.1.1 结合技术迭代、场景更新及行业伦理规范变化，定期开展伦理风险复盘评估，动态排查潜在风险隐患。

9.1.2 针对管控与运行中发现的问题，持续优化伦理风险识别、处置、管控机制，实现风险管理体系动态迭代。

9.1.3 结合行业案例与监管要求，适时更新伦理风险清单和管控策略，适配行业治理发展需求。

9.2 多方协同保障

9.2.1 建立研发、制造、运营、使用、监管多方协同治理机制，明确各方职责，形成共治共管的伦理风险治理格局。

9.2.2 健全常态化沟通反馈机制，畅通风险上报与整改协同渠道，高效处置各类伦理风险问题。

9.2.3 依托行业与科研平台开展伦理交流，共享治理经验，提升行业整体伦理风险管控能力。

9.3 伦理监督与审查

9.3.1 建立伦理合规审查机制，将伦理审查纳入研发、上线、迭代全流程，强化源头风险防控。

9.3.2 开展常态化监督与专项抽查，督导整改伦理管控薄弱问题，压实主体责任。

9.3.3 对高风险应用场景实施重点监督复核，防范重复性、系统性伦理风险。

9.4 能力建设与持续改进

9.4.1 面向全链条从业人员开展伦理风险防控培训，提升岗位伦理风险防控意识与履职能力。

9.4.2 普及智能机器人伦理准则与管控要求，强化各相关方向善、合规、可控的应用理念。

9.4.3 持续开展伦理治理能力建设，完善技术、制度、流程相结合的长效保障体系。

附 录 A
(资料性)
智能机器人伦理风险清单

A.1 人机交互伦理风险清单

人机交互伦理风险清单见表A.1。

表A.1 人机交互伦理风险清单

风险类型	发生阶段	风险项	风险点	描述
算法偏见与歧视	设计研发、部署应用、运行维护	视觉识别偏见	年龄识别偏差	机器人视觉算法对老年人、未成年人面部特征识别准确率低，易出现识别失败、服务拒绝、响应延迟等问题，造成年龄差异化服务不公
	设计研发、部署应用、运行维护		外貌特征歧视	算法对深色肤色、特殊妆容、疤痕、残障外貌群体识别容错率低，易出现误判、漏识别，产生隐性服务歧视
	设计研发、部署应用、运行维护	语音交互偏见	口音方言歧视	语音识别模型仅适配标准普通话，对方言、小众口音、外语口音识别率极低，导致非标准口音用户无法正常使用服务
	部署应用、运行维护	服务分配歧视	弱势群体资源排斥	公共服务、便民机器人优先响应青壮年用户需求，对老年人、残障人士服务响应滞后、资源分配不足，形成显性服务不公
隐私与数据	部署应用、运行维护	多模态数据过度采集	家庭环境无限制采集	家用、陪护机器人搭载高清摄像、收音设备，无间断采集室内环境、家具布局、家庭成员活动画面，无采集边界与时长限制
	部署应用、运行维护	生物特征数据滥用	人脸信息非授权比对	公共服务、安防机器人采集的人脸数据，超出服务必要范围，用于非授权身份比对、黑名单筛查、人员标记
	运行维护、回收处置	数据安全泄漏	本地存储数据泄露	在设备维修、报废、丢失过程中，未做数据清除处理，导致隐私信息被非法读取
权责界定与溯源	部署应用、运行维护	人机事故责任模糊	工业机械伤人责任不清	工业作业机器人碰撞、挤压、夹伤操作人员，因研发设计、设备制造、运维管控、现场使用多方因素交叉，无法明确唯一责任主体
	设计研发、运行维护	算法决策溯源困难	异常决策日志缺失	机器人关键决策无完整日志记录，出现误判、异常行为后，无法还原决策过程，难以定位问题根源与责任主体

表A.1 人机交互伦理风险清单（续）

风险类型	发生阶段	风险项	风险点	描述
权责界定与溯源	部署应用、运行维护	跨主体责任推诿	联合运营事故责任推诿	多方联合运营场景下（如设备厂商、运营方、使用单位），事故发生后各方相互推诿，合同条款责任划分模糊，无法形成追责闭环
算法黑箱与不可解释	设计研发、部署应用	关键决策不可解释	医疗决策逻辑不透明	医疗辅助机器人输出诊疗建议、康复方案、用药参考，无法解释决策依据与推理过程，医护人员无法核验有效性与安全
	设计研发、运行维护	算法漂移不可预判	模型迭代漂移失控	机器人算法在线迭代、持续学习过程中出现模型漂移，决策规则逐渐变化，无监测机制，风险不可预判
	部署应用、运行维护	自主行为突变无预警	导航路径异常跳变	机器人自主导航算法在环境变化时输出突变行为，如路径异常跳变、避障策略骤变，无前置预警机制，现场人员无法预判
恶意滥用与公共安全	运行维护、回收处置	实体篡改与破坏	程序破解恶意冲撞	不法人员破解机器人固件、篡改运行程序，操控机器人在公共场所冲撞行人、撞击公共设施，危害公共安全
	部署应用、运行维护	功能滥用与非法监控	无差别全域监控	巡逻、安防机器人被违规用于商圈、社区、校园无差别人脸采集、语音监听，实现大范围非法监控
	回收处置	退役设备违规改造	报废设备二次复用	报废设备未执行硬件销毁与程序清除，被非法回收后重新激活、改装，绕过安全管控投入违规使用
社会冲击与人伦异化	部署应用、运行维护	人机情感关系异化	独居老人情感依赖成瘾	养老陪护机器人过度模拟情感陪伴、共情回应，导致独居老人过度依赖机器人，弱化人际交往与亲情沟通
	部署应用、运行维护	社会就业与劳动冲击	基础岗位结构性失业	工业、物流、保洁、客服、餐饮服务机器人规模化落地，替代大量低技能基础岗位，引发区域性结构性失业问题
	部署应用、运行维护	人机身份认知混乱	人形机器人身份仿冒	人形机器人外观高度拟人化，在公共场景引发身份混淆，儿童、认知障碍群体难以区分人机身份，产生伦理认知偏差
基本权益与人身侵害	生产制造、运行维护	机械安全与人身伤害	工业机器人挤压夹伤	工业机器人关节失控、力控算法失效、安全防护失效，作业过程中挤压、夹伤、撞击现场操作人员

表A.1 人机交互伦理风险清单（续）

风险类型	发生阶段	风险项	风险点	描述
基本权益 与人身侵害	运行维护	特殊群体权益侵害	未成年人价值观误导	教育、陪伴机器人内容审核机制缺失，推送低俗、暴力、拜金等不良内容，误导未成年人价值观念养成
	部署应用、 运行维护	行动自由受限	公共空间通行阻断	公共服务机器人在执行任务时违规占用通道、限制人员通行自由，对残障人士、推婴儿车群体造成权益侵害

A.2 多机协同伦理风险清单

多机协同伦理风险清单见表A.2。

表A.2 多机协同伦理风险清单

风险类型	发生阶段	风险项	风险点	描述
协同合作	部署应用、 运行维护	任务分配不公	低价值任务集中化	多机器人协同任务分配算法未嵌入公平性约束，部分机器人持续承担低价值或高风险任务，形成“劳动不公”
	部署应用、 运行维护	资源调度偏差	资源垄断	资源调度算法偏袒高性能机器人，导致弱势机器人功能闲置，形成资源分配不公
	部署应用、 运行维护	协同决策边缘化	能力差异导致决策排斥	不同厂商机器人能力差异显著，协同决策中能力较弱的机器人意见被系统性忽略，形成“决策排斥”
竞争与冲突	部署应用、 运行维护	过度竞争	竞争危及人类安全	物流机器人在竞速配送中为抢占通行资源危及行人安全，竞争行为间接损害人类利益
	部署应用、 运行维护	合谋行为	排斥性联盟	多机器人在任务竞标中通过隐式通信形成价格合谋或联盟排斥，破坏公平竞争秩序
	部署应用、 运行维护	资源争用升级	通行资源死锁	多台机器人在有限通道资源下竞争，未设置统一仲裁机制，形成通行死锁，导致区域服务全面中断
数据交换 与互操作	部署应用、 运行维护	隐私跨系统扩散	共享数据未脱敏	机器人间共享数据中包含个人隐私信息未做脱敏处理，导致隐私信息随数据共享跨系统扩散
	设计研发、 部署应用	互操作协议不兼容	协同失效	不同厂商机器人通信协议不兼容，引发协同失效或安全事故

表A.2 多机协同伦理风险清单（续）

风险类型	发生阶段	风险项	风险点	描述
数据交换与互操作	部署应用、运行维护	权限越界获取	跨系统未授权操控	跨系统权限管理混乱，已授权的互操作权限未定期复核，导致未授权机器人获取敏感控制权限
级联风险与安全边界	运行维护	故障传播	连锁偏航	单台导航机器人定位失效导致编队连锁偏航，单点异常通过协同网络传播放大
	运行维护	恶意指令扩散	入侵蔓延	单台机器人被恶意入侵后通过协同通信链路向其他机器人扩散恶意指令
	运行维护	连锁停机扩散	多米诺式服务中断	协同任务中一台机器人故障引发“多米诺”式连锁停机，冗余设计不足导致大面积服务中断
集体行为与涌现效应	运行维护	群体过度聚集	公共空间挤占	群体路径规划中涌现出对特定区域的过度聚集，引发公共空间挤占
	运行维护	群体共谋	排斥人类监督	集体决策中涌现出排斥人类监督的“共谋”行为模式，偏离设计目标后因缺乏人工干预而持续恶化
	运行维护	群体行为偏离	集体目标漂移失控	大规模机器人协同行为偏离设计目标后，因缺乏“人在回路”的群体监督纠偏机制而持续恶化

附录 B
(资料性)
典型应用场景示例

B.1 政务应用场景

B.1.1 政务场景基础信息

政务场景基础信息见表B.1。

表B.1 政务场景基础信息

项目	内容
机器人实体类型	政务大厅咨询引导机器人、自助办事终端机器人、身份核验机器人、政务数智员工（智能问答系统、智能审批辅助系统）
部署场所	政务服务大厅、社区便民服务中心、不动产登记及税务社保办事窗口
服务人群	办事群众、企业经办人、老年及残障办事群体
核心硬件	移动底盘、人脸采集传感器、身份证读取模块、触控交互终端、本地业务存储模块、自然语言处理引擎、知识图谱模块、大语言模型推理服务
场景运行特征	采集公民身份、户籍、社保等高度敏感政务数据，政务服务公平性、透明度、可及性要求最高；政务大厅内咨询引导机器人与自助办事终端并存，存在机间数据交换与协同服务场景；政务数智员工通过线上政务平台提供7×24小时智能问答与审批辅助服务，其全生命周期管控重点与实体机器人存在差异，需关注算法决策透明度、自动化审批公平性及系统退役时数据彻底清除等环节

B.1.2 政务场景专属伦理管控重点

政务场景专属伦理管控重点见表B.2。

表B.2 政务场景专属伦理管控重点

伦理管控维度		具体管控要求
人机交互	算法偏见与歧视防控	政务咨询、排队调度算法不得因办事人身份、地域、资产状况差异化分配服务资源；统一各群体服务标准，保障政务服务公平可及
	隐私与数据保护	身份证、人脸等生物特征信息仅业务办理期间采集，办理完毕即时清除本地缓存；政务数据严禁无审批跨境传输，人脸核验数据不得留存复用
	权责界定与溯源	咨询引导、材料预审、身份核验全流程留存操作日志；清晰划分设备厂商与政务服务管理部门责任，办事纠纷可完整复盘溯源
	算法黑箱与可解释	办事流程推荐、材料补正提示同步展示判定规则与依据；模型迭代变更全部备案，办事群众可清晰了解办理进度与判定逻辑
	恶意滥用与公共安全防控	后台运维分级授权，核验数据导出双人复核；封堵数据泄露漏洞，防止采集公民身份信息用于电信诈骗等违法行为
	社会冲击与人伦异化防控	保留人工服务窗口，确保老年人、残障等数字弱势群体可选择人工办事渠道；避免政务服务完全机器化，防止加剧数字鸿沟

表B.2 政务场景专属伦理管控重点（续）

伦理管控维度		具体管控要求
人机交互	基本权益与人身侵害防护	交互流程适配老年、残障人群认知水平，提供语音引导、大字界面等无障碍功能，保障全体公民平等获取政务服务的权利
多机协同	数据交换与互操作管控	政务大厅咨询引导机器人与自助办事终端之间的数据交换限定为业务必要范围，禁止跨设备共享公民身份证、人脸等敏感信息； 设备间通信采用加密协议，留存完整数据交换日志

B.2 养老应用场景

B.2.1 养老场景基础信息

养老场景基础信息见表B.3。

表B.3 养老场景基础信息

项目	内容
机器人实体类型	居家自主移动陪护机器人、养老院陪护机器人、康复机械臂实体设备
部署场所	居民住宅、养老院、社区日间照料中心等
服务人群	高龄独居老人、失能失智、肢体残障老年群体
核心机械与传感硬件	自主行走底盘、人体感应传感器、机械搀扶执行臂、音视频采集硬件、本地存储单元
场景运行特征	24 小时居家近距离人机共处，持续采集生理与居家私密数据，情感交互深度渗透日常生活，隐私、人伦异化是核心伦理矛盾

B.2.2 养老场景专属伦理管控重点

养老场景专属伦理管控重点见表B.4。

表B.4 养老场景专属伦理管控重点

伦理管控维度		具体管控要求
人机交互	算法偏见与歧视防控	跌倒、语音识别扩充高龄、口齿不清老年样本，统一全老年群体预警判定标准，不因身体机能差异降低识别服务精度
	隐私与数据保护	摄像设备配备私密区域手动遮挡开关； 生理、居家影像本地加密存储，严格管控远程查看权限； 禁止无审批跨境传输老年生物特征数据
	权责界定与溯源	永久留存跌倒预警、人机交互、体征传感完整记录； 明确设备厂商、养老机构、家属三方责任，老人权益受损事故可完整复盘定位诱因
	算法黑箱与可解释	健康风险预警同步展示判定指标与阈值； 模型自适应迭代全程存档，家属可完整查阅算法推理逻辑，消除黑盒治理隐患
	恶意滥用与公共安全防控	远程访问启用双重身份验证； 底层交互程序锁定防篡改，杜绝通过设备实施监听、虚假养老诈骗等侵害行为
	社会冲击与人伦异化防控	硬件限定每日情感交互时长，系统主动引导线下亲属、护工陪伴，避免老人产生重度人机情感依赖、弱化现实亲情

表B.2 政务场景专属伦理管控重点（续）

伦理管控维度		具体管控要求
人机交互	基本权益与人身侵害防护	交互逻辑适配失智、高龄老人认知水平，简化操作流程，保障独居、残障老人平等监护与陪伴服务权益
多机协同	协同合作管控	养老院多台陪护机器人任务分配引入公平性约束，避免重体力照护任务持续集中于单台设备； 机间共享老人健康数据须严格执行脱敏处理，禁止跨设备扩散生物特征信息

B.3 金融应用场景

B.3.1 金融场景基础信息

金融场景基础信息见表B.5。

表B.5 金融场景基础信息

项目	内容
机器人实体类型	网点移动咨询机器人、自助核验实体终端
部署场所	商业银行线下网点、社区金融服务站
服务人群	普通储户、老年客户、小微企业现场经办人
核心机械与传感硬件	移动机身、人脸采集传感器、触控执行终端、本地业务存储模块
场景运行特征	线下采集财产、征信等高敏感金融数据，资产差异易诱发算法歧视，财产纠纷溯源伦理需求突出

B.3.2 金融场景专属伦理管控重点

金融场景专属伦理管控重点见表B.6。

表B.6 金融场景专属伦理管控重点

伦理管控维度		具体管控要求
人机交互	算法偏见与歧视防控	对老年人、小微企业统一完整风险告知标准，消除金融服务歧视
	隐私与数据保护	人脸、账户影像仅业务办理期间采集，本地缓存定期自动清理； 无审批禁止对外、跨境传输客户金融敏感信息
	权责界定与溯源	线下咨询、风险测评全程录音存档； 清晰划分设备厂商、银行运营责任，理财误导造成财产损失可完整追溯追责
	算法黑箱与可解释	信贷授信、产品推荐同步展示完整评分明细； 模型迭代变更全部备案，客户可清晰查阅判定依据
	恶意滥用与公共安全防控	后台运维分级授权，修改推介话术双人复核； 封堵数据导出漏洞，防止采集人脸、账户信息用于电信诈骗
	社会冲击与人伦异化防控	高风险业务强制跳转人工柜员复核； 保留线下服务窗口，避免老年储户完全依赖机器、忽视人工风险提示
	基本权益与人身侵害防护	交互流程适配老年客户认知，完整披露全部风险条款，保障老年人公平金融服务与财产知情权
多机协同	数据交换与互操作管控	网点咨询机器人与自助核验终端之间的数据交换应限定为业务必要范围，禁止跨设备共享客户人脸、账户等敏感信息； 设备间通信采用加密协议，留存完整数据交换日志

B.4 医疗应用场景

B.4.1 医疗场景基础信息

医疗场景基础信息见表B.7。

表B.7 医疗场景基础信息

项目	内容
机器人实体类型	门诊导诊移动机器人、康复机械臂、院内转运机器人、手术辅助机器人
部署场所	综合医院、康复疗养院、社区卫生服务中心
服务人群	普通患者、儿童、老年慢性病患者、残障康复人群
核心机械与传感硬件	运动执行底盘、肢体机械臂、生理传感模块、影像采集硬件
场景运行特征	实体设备直接作用人体，采集高度敏感医疗数据，诊疗公平、医疗隐私、生命权益伦理底线要求最高；院内多种机器人共处同一物理空间，转运机器人与导诊机器人存在路径竞争与数据交互

B.4.2 医疗场景专属伦理管控重点

医疗场景专属伦理管控重点见表B.8。

表B.8 医疗场景专属伦理管控重点

伦理管控维度		具体管控要求
人机交互	算法偏见与歧视防控	病灶、康复评估均衡纳入儿童、高龄、残障病患训练样本，统一全人群诊疗判定标准，避免特殊群体诊疗适配不足
	隐私与数据保护	病历、影像数据院内硬件闭环存储； 设备外送维修前清空本地医疗缓存，严格执行医疗数据跨境审批制度，杜绝病患隐私泄露
	权责界定与溯源	康复、影像识别、转运操作参数永久归档； 划分设备厂商、医疗机构责任边界，诊疗损伤事故可完整复盘溯源
	算法黑箱与可解释	AI诊断、康复方案同步标注循证医学依据； 模型迭代完整存档，医护、患者可完整查看判定推理链路
	恶意滥用与公共安全防控	诊疗参数修改实行双人医护授权；严控病历数据导出权限，防止医疗信息倒卖、篡改诊疗建议误导病患
	社会冲击与人伦异化防控	明确机器人仅为诊疗辅助工具，重大诊疗决策必须执业医师确认； 保障线下医患沟通，避免过度依赖机器弱化人文医疗
	基本权益与人身侵害防护	康复、导诊交互适配儿童、老年、残障人群生理认知特征，保障弱势群体平等诊疗康复权益
多机协同	竞争与冲突管控	院内转运机器人、导诊机器人在走廊通行中设置统一路径仲裁机制，患者转运任务享有最高优先级，禁止机器人在通行权竞争中延误危重患者转运
	数据交换与互操作管控	导诊机器人与转运机器人共享患者就诊信息时严格执行分级分类管理，共享数据不得包含完整病历影像； 跨系统互操作权限定期复核，杜绝未授权机器人获取患者敏感信息

B.5 仓储场景

B.5.1 仓储场景基础信息

仓储场景基础信息见表B.9。

表B.9 仓储场景基础信息

项目	内容
机器人实体类型	多品牌自主移动机器人（AMR）、分拣机械臂、自动叉车
部署场所	大型智能仓储中心、物流分拣枢纽
服务人群	仓储作业人员、物流调度员、现场管理员
核心硬件	激光雷达、视觉导航模块、无线通信模块、机械执行臂、载重底盘
场景运行特征	多品牌异构机器人在同一物理空间高密度协同作业，机间通信频繁，级联风险、资源竞争、涌现效应是核心伦理矛盾；仓储场景中机器人与作业人员混合通行，人机交互与多机协同风险叠加，需统筹管控

B.5.2 仓储专属伦理管控重点

仓储场景专属伦理管控重点见表B.10。

表B.10 仓储场景专属伦理管控重点

伦理管控维度		具体管控要求
人机交互	基本权益与人身侵害防护（人机）	保留人工调度优先干预权限，在高风险区域设置人工监督岗； 机器人异常行为实时告警推送至现场管理员，保障作业人员人身安全； 人机混行区域设置物理隔离与速度限制，防止机器人冲撞、挤压作业人员
多机协同	协同合作管控	任务分配算法引入公平性指标，避免低价值搬运任务持续集中于部分老旧机器人； 定期审计资源调度结果
	竞争与冲突管控	路径规划竞争中设置人类安全优先级最高规则； 部署统一仲裁机制解决通行权冲突； 禁止机器人通过隐式通信形成通行优先级联盟
	数据交换与互操作管控	机间共享数据严格执行分级分类管理，共享数据中不得包含人员个人信息； 跨品牌互操作采用统一安全协议； 数据交换日志完整留存
	级联风险与安全边界防控	部署单台机器人异常行为熔断机制； 协同通信引入消息完整性校验；定期开展级联失效压力测试； 设置系统级紧急停止按钮

表B.10 仓储场景专属伦理管控重点（续）

伦理管控维度		具体管控要求
多机 协同	集体行为与涌现效应 防控	部署前开展大规模群体行为仿真； 运行中部署群体行为偏差监测系统； 设置群体行为伦理边界阈值与人工纠偏机制

参 考 文 献

- [1] GB/T 41527—2022 家用和类似用途服务机器人安全通用要求
 - [2] GB/T 45502—2025 服务机器人信息安全通用要求
 - [3] DB37/T 4845—2025 人工智能技术应用伦理风险的治理要求
 - [4] 工业和信息化部人工智能标准化技术委员会. 工业和信息化领域人工智能安全治理标准体系建设指南（2025）
 - [5] 工业和信息化部. 工业和信息化部关于印发《人形机器人创新发展指导意见》的通知：工信部科〔2023〕193号. 2023年
 - [6] 工业和信息化部，国家发展和改革委员会，科学技术部等. 十五部门关于印发《“十四五”机器人产业发展规划》的通知. 工信部联规〔2021〕206号. 2021年
 - [7] 国家人工智能标准化总体组，全国信标委人工智能分委会. 人工智能伦理治理标准化指南. 2023年
 - [8] 中国国家新一代人工智能治理专业委员会. 新一代人工智能伦理规范. 2021年
-