

优 必 选
UBTECH



人形机器人科技伦理治理白皮书

(2026)

二〇二六年九月

致 谢

本白皮书在编制过程中承蒙以下人员及单位的热忱参与和辛勤工作，特此致谢。

指导专家

张 平 张 霄 杨庆峰

编审专家

邓 峰 肖少平 周 文 李兰兰 王丽娟

参编专家（按姓氏笔画排序）

刘 伟 李修林 袁 满 高 雅 黄增桃 梁乔玲

参编单位

深圳市优必选科技股份有限公司

深圳市律师协会

工业和信息化部电子第五研究所

深圳市区域科技伦理服务平台

深圳市标准技术研究院

版权声明

本白皮书版权属于深圳市优必选科技股份有限公司，并受法律保护。凡转载、摘编或援引本白皮书内容者，应注明来源为“深圳市优必选科技股份有限公司编著《人形机器人科技伦理治理白皮书》”，违者必究。

序 言

数字智能体和具身智能体的出现，标志着人类将从过去“人主动适应机器”走进了“机器主动适应人”的时代。随着人形机器人正逐步走出实验室和展厅，进入工厂、商场、学校、医疗机构和家庭，人形机器人将开始适应人类的生活规则和伦理要求，其部署条件的开放化、交互对象的公众化、人机关系的持续化等变化，导致现有的伦理治理范式在治理对象、风险管理、责任分配等层面均面临挑战。

本白皮书针对伦理、法律和安全的边界模糊，伦理风险识别和描述方法缺失，风险控制与风险机理的脱节，多主体的伦理责任界面不清等当前规范与实践中的突出问题展开。基于现行科技伦理治理框架对于人形机器人的具身性、拟人性和关系性风险属性覆盖不足的现实，本白皮书提出了行动可控、拟人适当、交互有度等三项人形机器人伦理专项保护目标，并构建了一套以人形机器人伦理治理为目标的七步风险管理框架，将抽象的人工智能伦理原则转化为可识别、可评估、可控制、可测试、可审查的具体落地要求，帮助不同专业背景的伦理风险评估和审查人员，建立统一事实基础与分析评价框架。

当然，具身智能技术仍处于快速发展阶段，人形机器人真正走进千家万户尚需数年，而且具身模型等核心的技术路径尚未收敛，其商业落地对个人、家庭、社会等关系层面的影响也缺乏长期、公开且可对比的研究数据。本白皮书本着适度前瞻、以终为始的思路，提出的诸多观点和设计仍属于探索性内容，难免存在考虑不周之处。我们愿与行业各方共同验证、修正和完善相关内容，也恳请广大读者批评指正。

目录

第一章 人形机器人科技伦理治理范围	1
1.1 人形机器人的定义	1
1.2 人形机器人科技伦理治理问题的提出	2
1.3 人形机器人科技伦理治理的相关概念	3
1.4 人形机器人科技伦理治理的边界	6
第二章 全球科技伦理治理框架概况	9
2.1 主要司法管辖区的科技伦理治理框架	10
2.2 国际组织及标准机构	16
2.3 现有治理框架对物理行动的覆盖情况	17
2.4 现有治理框架对拟人交互的覆盖情况	19
2.5 现有治理框架不足的总结	21
第三章 人形机器人伦理风险成因	25
3.1 人形机器人伦理风险的技术来源	25
3.2 人形机器人的特有伦理风险	28
3.3 人形机器人专项伦理保护目标	39
第四章 人形机器人伦理风险识别	43
4.1 人形机器人伦理风险的描述和识别	43
4.2 形成产品部署画像	44
4.3 确定受保护主体与伦理影响	47
4.4 构建风险形成机制	51
4.5 梳理伦理影响并补充评估标签	52
4.6 形成伦理风险描述表单	53
第五章 人形机器人伦理风险评估	55
5.1 伦理风险评估与其他专项评估的联系	55
5.2 人形机器人伦理风险的评估要素	57
5.3 伦理可接受条件	62

第六章 人形机器人伦理风险控制	63
6.1 人形机器人的伦理风险控制措施	63
6.2 人形机器人伦理风险技术测试与验证	69
6.3 人形机器人伦理审查	73
第七章 全生命周期动态伦理治理	75
7.1 全生命周期动态伦理治理的必要性	75
7.2 立项阶段	76
7.3 设计开发阶段	76
7.4 上市前审查阶段	77
7.5 产品部署阶段	79
7.6 产品上市后阶段	80
7.7 产品更新迭代阶段	81
7.8 停服阶段	82
第八章 人形机器人多主体伦理责任共治	87
8.1 制造商	87
8.2 供应商	88
8.3 部署方	89
8.4 运营商	89
8.5 二次开发者	90
8.6 经销商	90
8.7 用户	90
8.8 其他第三方	91
8.9 行业共治建议	92
附录 I 典型应用场景下的伦理风险分析	97
附录 II 全球涉人工智能与机器人领域科技伦理重要文件	109
参考文献	111

第一章 人形机器人科技伦理治理范围

1.1 人形机器人的定义

本白皮书所称人形机器人，是以人体整体结构、主要身体部位或整体拟人行动方式为基本形态，能够依靠感知、决策、规划和执行等能力在物理环境中自主或半自主完成任务，并可能通过语言、动作、表情或角色设定实现与人进行交互的智能实体。这一定义仅用于本白皮书对于人形机器人科技伦理治理的分析，不代表任何司法管辖区域内的法律法规或标准文件中对人形机器人相关产品、系统或设备的正式定义。其判定要素具体如下：

第一，具有可识别的人体整体结构、主要身体部位或整体拟人行动方式。完整人体比例、头躯干四肢结构、仿生面部以及整体拟人行动方式，均可作为认定机器人拟人特征的显著标志。人类自然语言、情绪反馈和角色设定等因素虽然能够增强拟人化效果，但应该主要用于评估风险暴露程度和治理强度，不宜独立作为人形机器人形态的认定依据。

第二，能够感知物理空间的环境信息，并通过自身执行单元实施动作并与周围现实环境发生交互。轮式移动、双足行走、抓取、递送和身体接触等均属于与现实环境交互的物理动作。人形机器人物理行动能力的强弱可能影响风险等级，但不是必然引发伦理风险的判断依据。陪伴人形机器人可能产生较强的情感关系影响，但物理行动能力弱，与周围物理环境的交互往往有限；工业人形机器人可能几乎不涉及情感交互，但是物理行动能力较强，能频繁、直接地与周围物理环境发生交互，可能带来更突出的物理安全风险。

第三，该实体具备人工智能和交互运行能力，能够在一定程度上适应交互场景或维持持续的互动关系。这些能力包括感知、生成、规划、学习、记忆用户偏好、主动发起联系、保持角色一致性，以及根据反馈调整行为等，并且与人的交互亦可呈现不同程度的社会性特征。

人形机器人无需同时具备以上所有要素，只要具备其中部分能力并因此产生相应的伦理影响，即可纳入本白皮书对人形机器人科技伦理治理的研究范围。

上述定义保持一定的开放性，因为人形机器人产业仍处于快速发展阶段，拟人外观、运动控制方式、具身模型架构和应用场景都在持续创新。不能仅凭人形机器人的名称，或者其双足行走、仿生面孔等产品特征来划定对人形机器人的治理边界，这样可能会导致对能力差异显著的人形机器人产品采取同样的治理强度，这也是当前治理实践中的常见问题。

另一个需要进一步辨析的问题，是具身智能、智能机器人、陪伴人形机器人等相邻概念。具身智能强调智能系统运行于物理世界之中，从环境中学习、推理并行动，其实体可以是多种形态，比如自动驾驶汽车、无人机等。人形机器人是具身智能的重要实现形式，但具身智能并不以拟人身体和社会表达能力为必要条件。智能机器人的涵盖范围更广，包括工业机械臂、移动平台和专用服务机器人等。陪伴人形机器人则是通过机器人的功能和角色定位来进行定义，陪伴人形机器人可以呈现人形、动物形或其他外观。

1.2 人形机器人科技伦理治理问题的提出

将人形机器人作为专门的科技伦理治理课题开展研究，并非仅仅是因为它拥有不同于一般工具型功能设备的拟人外观，关键在于与传统机器设备及非人形具身智能设备相比，人形机器人在部署环境、交互对象和使用方式等方面均发生了显著变化，并对既有的科技伦理治理框架提出了新的挑战，主要表现在以下几个方面。

第一，人形机器人的部署情境开始从受控或封闭的专用空间转向人类活动的开放空间。工业机器人、智能家电等传统功能型设备，以及无人驾驶、无人机等新兴具身智能设备，大多仍部署在机器专用的特定环境中，使用者需要掌握特定的操控技能才能完成相应任务。随着人类社会迈入人机共生阶段，人形机器人最有可能作为新的物理智能体开始走进人类生活，将适应人类的心理空间、生理空间和社会空间，并参与人类生活的相关规则。这意味着机器必须理解人类的语言，并适应人类世界的物理运行规则、法律约束与伦理道德准则。过去基于“人适应机器”逻辑制定的安全和伦理治理框架，在“机器适应人”的新时代情境下有必要重新审视。

第二，交互对象从专业操作者扩展到普通公众。部署在受控或有限空间内的机器设备，交互对象通常限于经过培训的专业操作者。人形机器人的交互对象范围更广、构成更复杂，家庭成员、顾客、访客、学生等各类身份群体均可能成为交互对象，其中相当一部分人可能根本没有主动选择过与机器人互动，甚至也没有意识到自己正在和机器人互动，更未曾像注册用户那样提前阅读产品说明，也缺乏行使相关个人权利的渠道。在此情形下，过去以用户知情同意为核心的合同式合规治理机制，可能难以直接适用。

第三，使用方式从单次任务演变为持续的交互关系。传统机器设备一般以完成单一特定任务为目标，任务结束即人机交互终止。但在人机共生场景下，具备长期记忆、稳定人格表现和主动交互能力的人形机器人，其影响不再局限于单次使用。用户可能在认知、情感等层面形成对机器人的期待与依赖。产品的系统更新、人格调整、记忆迁移、能力变化乃至服务中断或终止，均可能改变用户的行为习惯与心理预期。基于

单次交互和单一维度建立的传统风险评估方法，难以有效准确识别、评估和覆盖这种在时间、心理与行为等多个维度上逐步积累的复合型风险。

1.3 人形机器人科技伦理治理的相关概念

人形机器人同时涵盖人工智能技术与机器人技术，两种技术在人形机器人上深度耦合，决定了其科技伦理治理既需要综合考量二者的叠加影响，也需要分别识别各自特有的风险维度。如果不对相关概念加以区分，很容易导致治理对象识别模糊、风险分类失准，进而引发法律法规与伦理审查机制适用不当的问题。

1.3.1 伦理

“伦理”，在中国传统语境下，指处理人与人之间关系的道理和原则。在西方哲学史上，伦理学（Ethics）又称道德哲学（Moral Philosophy），是哲学的一门分支学科，核心研究主题是道德价值，旨在从理论层面构建指导个体行为的法则体系，探究道德的起源、本质、作用及发展规律。

随着科学技术的快速发展，技术已逐渐从人类的辅助工具，转变为深刻影响社会结构、经济运行与个人生活的核心力量。与此同时，科技创新活动本身也催生了大量新兴伦理问题。例如，人工智能系统可能引发的算法歧视与自动化决策责任问题，基因编辑技术涉及的人类生命边界问题，以及机器人进入家庭、医疗和公共服务场景后带来的人机关系变化问题，这些问题均已超出了传统技术评价体系的覆盖范围。因此，伦理学的“拷问”逐步向科技领域延伸，最终形成科技伦理这一重要的应用伦理学分支。

1.3.2 科技伦理

联合国教科文组织将科技伦理定义为“指导科学研究与技术创新及应用过程的道德原则”，2022年3月中共中央办公厅、国务院办公厅印发的《关于加强科技伦理治理的意见》中将科技伦理界定为“开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范”。

上述定义表明，科技伦理兼具价值属性与规则属性，它既是一套价值理念或道德原则，能够为各类科技活动指明发展方向；又是一套行为规范，能够为具体科技实践提供可操作的约束框架。这种双重属性，决定了科技伦理治理既需要在价值层面开展教育引导与文化塑造，也需要在制度层面落实审查监管与规范约束。

科技伦理在价值维度和规则维度上的双重属性，使其能够覆盖科学研究、技术开发、成果转化和技术应用的全流程，其核心关注的是科技活动开展的正当性，具体而

言：一是科学研究层面，包含科研伦理，即科技人员开展科学研究过程中应当遵循的伦理规范；二是技术开发与应用层面，包含技术伦理、工程伦理等，即工程技术在开发应用、设计制造全流程中应当遵循的伦理规范；前者关注科研活动本身的正当性，后者关注技术应用及其后果的正当性，二者在治理上也往往对应不同的管理主体与审查机制。

1.3.3 人工智能伦理

联合国教科文组织 2021 年发布的《人工智能伦理问题建议书》（Recommendation on the Ethics of Artificial Intelligence）指出：“人工智能伦理是对人工智能技术开展规范性评估与指导的动态基础，它以人的尊严、福祉以及防范损害为导向，植根于科技伦理的框架之中。”

人工智能伦理是规范人工智能技术发展的价值体系、伦理原则与行为框架，以维护人类尊严、增进人类福祉、防范技术伤害为核心要求，旨在引导社会以负责任的态度应对人工智能技术对人类个体、社会运行及生态环境的影响，并为社会接纳或拒绝特定人工智能应用提供伦理层面的判断依据。

从研究对象和全球人工智能治理的普遍实践来看，虽然不同国际组织或国家的具体表述可能略有差异，国际公认的核心人工智能伦理原则可以归纳为六大基础原则：公平与公正（Fairness & Justice）、透明与可解释性（Transparency & Explainability）、安全与稳健（Robustness, Security & Safety）、隐私与数据保护（Privacy & Data Protection）、责任与问责（Accountability & Responsibility）、以人为本与人类福祉（Human-Centered Values & Well-being）。人工智能伦理关注的本质是人类如何以符合伦理要求的方式设计和使用人工智能技术。

1.3.4 机器人伦理

机器人伦理学研究人类在开发、部署与使用机器人的过程中应当遵守的道德规范。其核心关注点是“关于机器人的伦理”（Ethics for Robots），即通过伦理原则约束机器人设计者与使用者的行为，而非认定机器人本身已成为具备道德责任能力的主体。

与人工智能系统主要存在于数字环境中不同，机器人拥有实体形态，能够通过传感器、执行单元和自主控制系统对现实环境产生实际影响。随着机器人逐步从工业制造领域延伸至家庭、医疗、教育和公共服务等应用场景，机器人伦理研究的问题也在持续拓展。一方面，需要保障机器人运行中的物理安全与可靠控制，避免其自主行为对人体和环境造成损害；另一方面，也需要关注机器人高度拟人化后，可能引发的人类情感依赖、社会关系替代、人与机器边界模糊等问题。

对于人形机器人而言，它兼具人工智能能力、实体行动能力与拟人交互能力。因

此，既有人工智能伦理治理框架已无法为其提供足够的规范支撑，仍需从机器人伦理的视角出发，考量具身性与社会交互带来的特殊风险。这正是将人形机器人作为独立科技伦理治理课题的关键依据。

1.3.5 机器伦理

目前学界对其探讨主要停留在学术层面，尚未形成统一的官方定义。部分中国学者认为，机器伦理研究如何在计算机、机器人或其他自动机器中设计、实现并生成符合规范的伦理决策，可以以研究人类思考、学习以及作出道德决策的过程为参考，探索将伦理规范嵌入机器决策的路径。部分美国学者提出，机器伦理的核心目标是研究如何设计出具备独立伦理决策能力的软件系统。综上，机器伦理关注的是如何使机器系统获得伦理判断能力，使其在自主运行过程中能够采取符合人类伦理要求的行为。

虽然机器人伦理与机器伦理仅一字之差，二者的范畴却相差极大。机器伦理聚焦于机器本身应当嵌入何种伦理决策，例如当机器人看到主人遭到坏人攻击时，它应当优先保护主人避免对攻击者造成伤害还是协助主人反击，这类问题就可能涉及机器的道德主体资格、伦理规则形式化、机器自主决策以及责任归属等基础理论问题。目前，机器伦理仍处于探索阶段，尽管自动驾驶、智能机器人等领域已经开始尝试将伦理规则嵌入系统决策，但围绕机器能否承担道德责任，学界仍未形成统一观点。

1.3.6 相关概念比较

上述概念虽同属于伦理治理范畴，但其关注对象与治理重点存在明显差异。科技伦理、人工智能伦理和机器人伦理，均属于对人类技术活动进行规范的伦理领域，而机器伦理则更进一步，探索机器是否能够成为独立的伦理主体。

表 1-1 相关概念比较

概念	主要规范对象	核心问题
伦理	人类行为	什么行为具有道德正当性
科技伦理	科技活动主体	技术是否应被开发及如何发展
人工智能伦理	人工智能研发、部署和使用主体	如何保证人工智能安全、公平、透明
机器人伦理	机器人设计和使用者主体	如何保证机器人在人类环境中的安全 互动
机器伦理	机器自主决策系统	机器是否能够进行伦理判断

科技伦理、人工智能伦理、机器人伦理与机器伦理四者之间既有联系也有区别。

其中科技伦理是人工智能伦理、机器人伦理等细分领域的上位概念，为各类科技创新活动提供了具有普遍性的价值框架。人工智能伦理和机器人伦理的研究对象各有侧重，人工智能伦理聚焦数字世界中的信息、内容与决策问题，机器人伦理则关注物理世界中的人机交互问题。机器伦理试图突破“人是唯一道德主体”的传统界定，探索机器参与道德决策的可能性，并在此基础上甚至可能催生出全新的伦理关系。

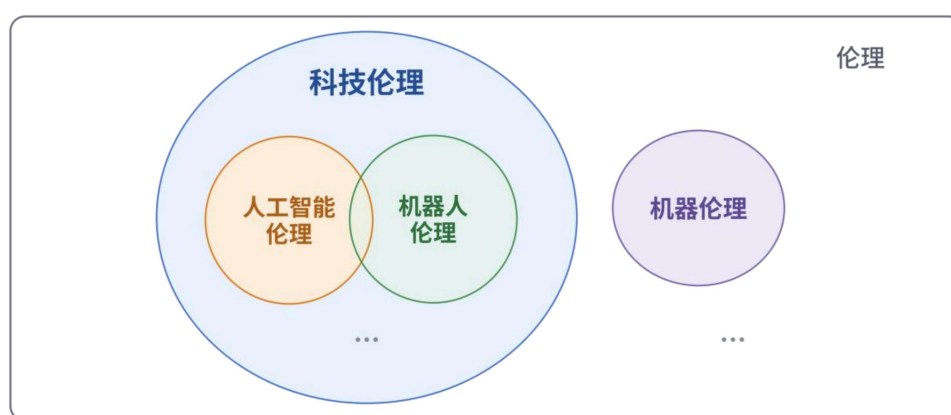


图 1-1 相关概念关系

上述差异意味着，人形机器人的伦理治理既无法单独套用人工智能伦理的既有框架，也无法完全移植传统机器人伦理的框架。若不明确界定治理边界，就可能将伦理讨论泛化为对所有问题的无限延伸，或将其与法律合规、技术安全混为一谈。

1.4 人形机器人科技伦理治理的边界

1.4.1 与法律、安全的边界

人形机器人科技伦理治理的另一个突出问题是当前治理边界并不清晰，这会导致企业在开展治理工作时，法律、安全和伦理三者的治理边界可能出现交叉重叠。保护生命健康、个人信息、消费者权益与公平机会既可能是法律层面的明确要求，也可能是安全领域的核心目标，或是伦理范畴的重点关切。例如某一台人形机器人发生的单次碰撞事件，就有可能同时涉及产品缺陷、机械安全与伦理三类问题。因此，厘清三者的核心区别，对明确人形机器人科技伦理治理的边界与核心关注事项至关重要，有助于避免伦理体系因边界模糊而不聚焦。

事实上，法律、安全和伦理三类治理所依据的规范、聚焦的问题以及结论的效力各不相同。法律以特定司法管辖区的现行法律法规、监管解释和适用事实为依据，确定责任主体负有哪些法定义务，哪些行为被禁止或被要求，以及违反规则产生哪些法律后果，其效力来源于立法，具有强制性与可执行性。当人形机器人进行跨境销售或

进入医疗、教育、劳动等应用场景时，适用规则会因法域、主体、用途和产品分类的不同而发生变化。安全关注危险源、故障、性能边界和风险控制，关注的是人形机器人能否在规定条件下可靠运行，伤害风险是否被降至可接受水平。机械安全、电气安全、功能安全、网络安全和人工智能安全，共同构成了人形机器人安全和伦理治理的重要基础，其判断依据一般是可测量的指标和可验证的方法，结论具有可复现性。伦理则关注科技活动及其制度安排是否具有正当性，它会进一步关注某项功能是否值得开发、风险和收益由谁承担、用户能否理解并拒绝相关功能、某种拟人设计是否利用了人的脆弱处境，以及在法律尚未作出明确规定时，制造商应当如何作出负责任选择。其判断需要明确说明受影响的价值、相关利益主体、可能产生的后果、可替代方案和支撑证据。

在三类治理之下，还需进一步区分法律责任、伦理责任与行为责任。法律责任由有权机关依据具体法域的法律规范认定，后果包括行政处罚、民事赔偿、合同违约救济等，其成立需要同时满足主体身份、行为、损害、因果关系等法定构成要件。伦理责任由行业共识、企业内部治理机制和专业规范确定，后果通常表现为审查不通过、功能受限、上市延迟或整改要求，它依据伦理原则和风险评估结论作出判断，不要求损害已经实际发生。行为责任则是指主体在特定关系中应当采取或避免某类行为的要求，其来源可能是法律，也可能仅来自伦理或职业规范。由此可以明确三条边界：

第一，不得以合法替代伦理。合规不等于合乎伦理，产品即使符合所有适用法律，仍可能存在不可接受的伦理风险。以“法无禁止”作为伦理审查作为反驳伦理审查不通过结论的理由，实质上相当于否定了伦理治理的作用。

第二，不得以伦理替代法律。企业通过内部伦理审查不减免其备案、报告、记录保存和用户权利响应等法定义务。伦理框架不能作为解释法律义务范围的依据，审查通过也不构成法律免责。审查结论仅表明在给定条件下未发现不可接受风险，而不代表产品不会造成损害。反过来，伦理审查不通过也不能视为违法，审查可以基于比法律更高的标准作出否定判断，比如认定某项拟人化设计存在不可接受的依赖诱导风险，即使该设计在所有适用法域中均无禁止性规定。

第三，不得以用户的行为规范替代企业的结构性义务。要求用户不得滥用、要求操作者遵守规程，不能替代产品本身应具备的技术防护能力。

1.4.2 本白皮书的讨论边界

本白皮书的讨论范围将人类及其组织明确为伦理的责任主体，这一立场是为了与机器伦理有关的讨论划定边界。人形机器人虽然具备行动能力，技术上能够作出自主决策和行动，其行为也需要进行伦理影响评估，但这些事实在目前的技术发展阶段还无法直接视为机器人拥有和人类相同的情感、能够理解道德依据、自主作出道德决策

或承受道德责难。因此，本白皮书不针对机器人权利、电子人格、机器意识或机器的道德主体资格展开讨论，凡涉及伦理判断与责任归属，仍以人类及其组织作为责任主体。

第二章 全球科技伦理治理框架概况

全球不同法域的科技伦理治理在核心议题与发展阶段上均存在差异，同一治理要求可能以法律、政策、标准、审查程序等多种规范形态呈现，其在效力层级、调整范围与实施条件方面存在显著差异。为构建统一的比较分析框架，本章将现有科技伦理治理工具划分为六类。

第一类是具有约束力的法律法规、规章以及规范性文件。数据保护、网络安全、产品安全、人工智能、消费者保护和产品责任等领域的相关制度，划定了市场主体必须遵守的合规底线，也是开展监管与救济工作的主要依据。这类治理工具法律效力强，但制定和修订的周期通常较长，当一项技术同时涉及多个制度领域时，还可能出现适用条件不清晰、主管部门权责划分不一致的问题。

第二类是国家层面的政策、伦理原则与业务指南。这类规则能够快速明确人工智能的治理方向，为各类参与主体提供行为参考。日本出台的《以人为中心的人工智能社会原则》《AI 企业指南》，以及美国发布的《人工智能权利法案蓝图》，都属于这一类别。这类指南可以伴随技术的发展及时完成更新，但往往需要依靠企业自愿采纳、政府采购引导或部门监管推动，才能真正落地见效。

第三类是风险管理与组织管理类标准，比较典型的包括 NIST AI RMF 1.0，即美国国家标准与技术研究院人工智能风险管理框架 1.0 版，以及 ISO/IEC 42001 人工智能管理体系、ISO/IEC 23894 人工智能风险管理指南、ISO 31000 风险管理指南等。这类标准为各类组织提供了统一的术语规范、实施流程与证据框架，既方便在跨行业场景中复用，也有助于将人工智能伦理原则融入现有管理体系之中。不过，这类标准也存在局限性：标准的通用性越强，就越需要针对具体产品特有的物理动作、关系角色与应用场景，另行补充适配性要求。

第四类是产品与行业技术标准。机械安全、功能安全、机器人安全以及人形机器人伦理等相关标准，可将相关要求落到设计、测试和验收的具体环节中。这类标准贴近工程实践，通常会比专门立法更早出台；不过不同标准在适用范围、强制程度和引用关系上并不统一，仍需要企业自行判断该如何组合适用。

第五类是测试、影响评估与伦理审查工具。AI Verify、人工智能影响评估、科技伦理审查等机制将治理关口前移，推动相关组织在开展 AI 研发、部署或进行重大变更前提前排查风险。这类工具的治理效果取决于评价对象覆盖是否完整、评估方法是否适配对应产品，以及审查结论能否切实影响研发与运营决策。

第六类是试点、沙盒和先导机制。当新技术的收益与风险尚未明朗时，在有限场景内开展试点运行，能够积累实证经验，为后续的标准修订与法律调整提供现实依据。

人形机器人一旦进入家庭、学校、医院、商场和工业现场，在真实运行过程中会出现各类实验室无法穷尽的行为组合，试点能够帮助治理者提前感知并厘清这些复杂状况，但试点本身无法免除事后的责任追究、权利救济与持续监督。

这六类工具各有侧重、相互补充。法律划定治理底线与责任归属，标准和指南明确实施路径，测试与审查生成合规证据，沙盒试点则检验规则在真实场景中的落地效果。判断某一工具是否适用的关键在于，它能否有效识别人形机器人带来的增量风险，并在此基础上构建可执行、可验证、可救济的治理框架。

2.1 主要司法管辖区的科技伦理治理框架

2.1.1 欧盟

欧盟采取基本权利、产品安全与市场准入的组合规制框架，已形成覆盖数据、产品安全、人工智能及损害救济等多个领域，并以政策与伦理原则为引导的多层治理框架。

在数据层面，2016年4月，欧盟通过的《通用数据保护条例》（General Data Protection Regulation, GDPR）对个人数据处理的合法性、透明度、目的限制、数据最小化和数据主体权利作出了系统规定，其第二十二条还就完全基于自动化处理且对个人产生法律效果或类似重大影响的决策作出了专门规定。人形机器人在处理家庭空间、面部、声音、位置和健康等数据时，GDPR 为此类数据处理提供了完备的规范框架。

在产品安全层面，2023年6月，欧盟通过《机械法规》（Regulation (EU) 2023/1230 on machinery）。相较于原《机械指令》（Machinery Directive 2006/42/EC），此次修订扩大了机械产品范围，将与机器安全相关的软件和自主行为纳入规制范围，并把网络攻击可能影响机械安全的情形纳入基本健康与安全要求的覆盖范围。人形机器人因具备机械结构和执行机构，通常需要根据其具体用途和产品构成，纳入机械产品安全、风险评估与合格评定框架。

在人工智能层面，2024年7月，欧盟发布《人工智能法案》（Regulation (EU) 2024/1689 Artificial Intelligence Act, AI Act），该法案第一百一十三条规定了分阶段适用安排，其中禁止性规则和人工智能素养要求自2025年2月2日起适用。法案以禁止实践、高风险系统和特定透明义务等安排实施风险分级；对高风险人工智能系统，该法在风险管理、数据治理、技术文档、日志记录、人类监督以及准确性、稳健性、网络安全等方面提出系统性要求，并要求提供者建立上市后监测与严重事件报告机制。该法案一定程度上能够覆盖人形机器人的部分人工智能治理需求，机器人在部分应用场景中有可能进入到高风险产品的规制范围。其中，第5条关于禁止操纵、欺骗和利用年龄、残障或特定社会经济处境所致脆弱性的规定，为拟人交互设定了部分边界；

第 50 条的透明义务则有助于处理人与人工智能系统互动时的身份识别问题。

在损害救济层面，2024 年 10 月，欧盟修订通过《产品责任指令》（Product Liability Directive, PLD, Directive (EU) 2024/2853），将软件纳入产品概念，并处理相关服务、软件更新和网络安全脆弱性等数字产品特征，为人工智能与机器人造成损害后的民事救济提供了更适应数字产品的制度基础。

在政策与伦理层面，2017 年 2 月，欧洲议会通过《关于机器人民事法律规则的决议》（Civil Law Rules on Robotics）。该决议虽无法律约束力，但较早讨论了自主机器人、责任、保险、伦理准则及社会影响等问题，表明机器人很早就进入了欧盟的制度视野。2019 年 4 月，欧盟委员会高级别人工智能专家组发布《可信人工智能伦理指南》（Ethics Guidelines for Trustworthy AI），进一步建立了可信人工智能的价值与要求框架。

总体来看，欧盟框架的特点是多项组合工具相互衔接：GDPR 针对个人数据处理，《机械法规》针对产品与物理安全，AI Act 针对人工智能风险，PLD 针对缺陷产品损害。但前述工具的边界需通过具体产品和协调标准进一步落实。《机械法规》和 AI Act 各自设有风险管理要求，两者如何共同验证人形机器人感知、模型判断、机械执行构成的完整行为链条，目前尚无统一答案，这也是欧盟规制组合后续需要解决的问题。

2.1.2 美国

美国联邦层面的人工智能治理长期以行政命令、部门既有执法权限和技术标准为主要规制手段。2026 年 3 月，白宫发布《国家人工智能政策框架》（A National Policy Framework for Artificial Intelligence）；同年 6 月，美国总统签署第 14409 号行政令《促进先进人工智能创新与安全》（Promoting Advanced Artificial Intelligence Innovation and Security），要求联邦机构围绕前沿模型的基准评估与自愿性发布前评估、网络安全漏洞协调以及关键基础设施防护作出部署。

2023 年 1 月，美国国家标准与技术研究院（NIST）正式发布《人工智能风险管理框架》（AI Risk Management Framework, AI RMF 1.0），以治理、映射、测量和管理四项功能帮助组织识别和应对人工智能风险，适用于不同部门和用途，强调有效可靠、安全、韧性、可问责与透明、可解释、隐私增强以及公平等可信特征。美国联邦层面仍在尝试以政策和行政工具应对快速变化的人工智能风险，但其适用范围、与既有行政命令的衔接可能快速调整，其效力也可能被后续行政令撤销或取代，应注意以现行正式文本为准。

美国这套工具组合的优势在于迭代速度快：NIST AI 风险管理框架（AI RMF）能够帮助企业建立统一的风险沟通语言，行业监管者也可依托消费者保护、反歧视、产品安全相关规范及各部门法处理具体问题。人形机器人企业可以借助 AI RMF 梳理模

型与组织层面的风险，同时仍需遵守现行适用的产品、劳动、隐私及行业相关规则。但这一路径缺乏面向人形机器人的统一制度入口，通用风险框架无法自动转化为人形机器人的安全与合规测试要求，此外，美国各州规则与联邦政策并行的格局，进一步拉大了人形机器人跨地区部署的合规差异，举例来说，同一产品在不同州可能面临截然不同的隐私、生物识别和自动化决策监管要求。整体来看，美国现有工具体系能够快速响应新出现的风险，但在专门对象定义、持续关系治理和跨域责任衔接方面，规则体系仍较为分散。

2.1.3 日本

2018年12月，日本内阁府发布《以人为中心的人工智能社会原则》，以尊重人的尊严、多样性与包容性、可持续性为基本理念；同年6月，日本内阁府发布《AI战略2019》（AI Strategy 2019），将伦理原则嵌入“Society 5.0”国家构想。2025年5月，日本国会通过《人工智能相关技术研究开发及应用推进法》，进一步明确了国家推进人工智能研发与应用的基本方向和政府责任，体现了促进创新的制度取向。

日本经济产业省（METI）等部门持续更新相关指南。2026年3月，《AI企业指南》（AI Guidelines for Business Ver 1.2）面向开发者、提供者和业务使用者，分别提出风险识别、信息提供、数据质量、透明、问责与人类参与等要求。这种按角色划分的指南便于企业将共同原则融入业务流程，也能随模型和应用的变化较快更新；其角色划分思路与本白皮书第八章的主体配置有相似之处，但指南不涉及各个主体之间证据责任的分担与衔接问题。

日本法域治理路径的优势在于，它能够有效应对技术发展带来的不确定性，同时为相关参与主体预留了与责任比例相适配的治理空间，人形机器人的开发者、提供者与使用者可依照这一路径分担相应治理责任。但该路径对自愿采纳意愿、行业自身能力以及组织内部治理质量的依赖度较高，当人形机器人涉及近身物理交互或对长期社会关系产生影响时，这类通用业务指南仍需和产品安全、消费者保护、劳动规范等具体行业规则配合适用才能发挥作用。现有指南虽提示需关注人的尊严与潜在风险，但还未建立物理执行全流程的整机评价方法，也未能适配长期人机关系的专门救济机制。

2.1.4 韩国

韩国较早通过立法推动机器人产业发展。2008年3月，韩国国会通过《智能机器人开发与普及促进法》，以产业规划、扶持措施和基础设施建设促进智能机器人发展。依据该法及相关政策讨论形成的《机器人伦理宪章》目前仍为草案，内容涉及人类尊严、公共利益、透明、可控、责任、安全与信息安全等。

2020年12月，韩国科学技术信息通信部（MSIT）正式发布《国家人工智能伦理

标准》（National AI Ethics Standards），提出人的尊严、社会公共利益和技术合乎目的三项基本原则。2024 年 12 月，韩国国会正式通过《人工智能发展和信任基础构建基本法》（Basic Act on the Development of AI and the Building of Trust），确立了国家人工智能治理架构，并针对高风险人工智能、生成式人工智能、透明和安全等设置了差异化要求，该法自 2026 年 1 月起施行。

韩国经验的重要意义在于，机器人早已成为韩国明确的政策与法律规制对象，以此提供了一个产业促进规则、伦理原则到人工智能基本法逐步叠加完善的典型样本。但韩国早期机器人立法的核心任务主要是推动机器人技术开发与产业普及，并未针对人形机器人的拟人化表达、长期用户记忆以及社会关系替代等特性构建完整的合规义务体系；同时，人工智能基本法与现有机器人法之间，仍需要针对具体产品与应用场景的专门规则加以衔接。

2.1.5 新加坡

2019 年 1 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）发布《人工智能治理框架模型（1.0）》（Model AI Governance Framework），提出了透明性、公平性、人为参与、内部治理和利益相关方沟通等五大核心原则。次年，《人工智能治理框架模型（2.0）》发布，新版框架新增了可解释性维度，并配套发布了《实施指南》和典型案例汇编，将抽象原则具体化为企业可操作的管理流程与技术措施。

2022 年 5 月，IMDA 推出 AI Verify 测试框架及工具包。AI Verify 是一个面向企业的 AI 治理测试平台，旨在帮助机构通过标准化测试与过程校验的方式，客观验证其 AI 系统的性能表现、透明度与公平性，从而以技术手段证明其“负责任 AI”的合规能力。

2026 年 1 月，IMDA 发布《智能体人工智能治理示范框架》（Model AI Governance Framework for Agentic AI），将治理视野扩展到能够自主决策和采取行动的智能体人工智能。框架强调部署前风险评估、明确责任边界、限制系统权限，并在资金处理、敏感数据修改等高风险或不可逆操作中保留人工审批。这三类工具相互衔接，形成了指南明确要求、测试验证成效、智能体框架规范行动权限的完整工具包。

这一工具包实现了从治理“输出”到治理“行动”的转向。传统人工智能治理长期以信息处理为核心，关注的重点包括训练数据是否合规、模型输出是否公平、自动化决策能否解释、生成内容是否真实可识别；智能体框架将治理的关注点转向权限管控、行为边界、不可逆操作与人工审批环节，这既是治理对象从算法系统向行动型智能体拓展的明确信号，也让这类方案成为现有工具中与人形机器人适配性最高的一类。

但软件智能体与人形机器人这类硬件智能体仍然存在较大差异：软件智能体可以通过最低权限原则限制系统可调用的工具范围，限制付款、修改记录等操作的执行，

却无法解决机械臂接触人体的力度控制、感知误差向执行机构的传递、旁观者如何理解机器人动作意图这类硬件层面的问题。

2.1.6 中国

中国的科技伦理与人工智能治理已形成多层级监管体系。2019年6月，科技部发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作和敏捷治理八项原则。2021年9月，科技部进一步发布《新一代人工智能伦理规范》，将伦理要求延伸至管理、研发、供应和使用等各个环节。2021年12月，全国人大常委会修订《中华人民共和国科学技术进步法》，首次在法律层面明确提出完善“科技伦理制度规范”。2022年3月20日，中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，正式明确了科技伦理的五项原则，即：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明，并对伦理审查、监督管理、教育宣传和多方共治等提出要求。2023年的《科技伦理审查办法（试行）》，继续沿用前述五项原则的科技伦理表述，同时规定了科技伦理审查的主体、范围、程序、复核、监督与登记要求，将人工智能纳入需要重点关注的科技活动领域。2026年3月，工信部等十部门发布《人工智能科技伦理审查与服务办法（试行）》，进一步针对人工智能活动细化了审查和服务机制，并推动相关机构提供审查与咨询服务，并将人工智能科技伦理的原则明确为七项，即：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明、保护隐私安全、确保可控可信。

在互联网信息及人工智能领域，已形成针对具体技术和服务的专项规则。2022年11月，网信办、工信部等联合发布《互联网信息服务深度合成管理规定》，对数据和技术管理、内容标识、用户管理等作出要求。2023年7月，网信办、发改委等发布《生成式人工智能服务管理暂行办法》，规定了生成式人工智能服务提供者在内容、数据、个人信息和服务管理方面的义务。2025年10月，全国人大常委会修订了《中华人民共和国网络安全法》，新增“人工智能伦理规范”条款，将人工智能伦理要求直接嵌入到网络信息安全法律框架。

2026年4月，国家网信办、工信部等发布《人工智能拟人化互动服务管理暂行办法》，对拟人化互动服务中的身份提示、情感引导、未成年人保护和用户权益等提出了要求，将拟人化互动服务纳入专项治理。该办法与人形机器人高度相关，因为拟人表达、情绪反馈和角色设定本就是人形机器人交互的核心特征，该办法也是对人形机器人拟人性风险覆盖的直接规制工具。不过它的直接规制对象仍以拟人化互动服务为主。当机器人同时具备机械身体、现实行动能力和近身接触属性时，还需额外适用产品安全、机器人专用标准以及部署场景等领域的相关规则。

产业政策和标准化工作也开始向人形机器人领域延伸。工信部于 2023 年发布《人形机器人创新发展指导意见》，对整机、关键零部件、软件创新和应用生态作出部署。全国机器人标准化技术委员会（TC591）及其人形机器人分技术委员会（SC5）已将伦理纳入标准化工作，目前分别处于立项、征求意见等阶段。此外，《工业机器人生命周期风险评价方法》（GB/T 38642-2020）、《服务机器人功能安全评估》（GB/T 38260-2019）等现有标准，也为机械与机器人风险评估提供了基础。

2026 年 7 月 17 日，在上海举行的 2026 世界人工智能大会暨人工智能全球治理高级别会议上，工信部会同有关方面共同指导发布《国际人工智能伦理治理行动计划》，为全球人工智能治理贡献了兼具原则性与可操作的行动框架。该行动计划是中国面向国际社会提供的重要公共产品，在联合国《未来契约》及其附件《全球数字契约》的框架下，围绕人工智能全生命周期伦理治理、风险分类分级防控、敏捷治理机制建设、产业链协同生态培育、社会氛围营造等方向，呼吁各国结合自身实际积极行动，本着共商共建共享的原则加强政策协调与实践合作。

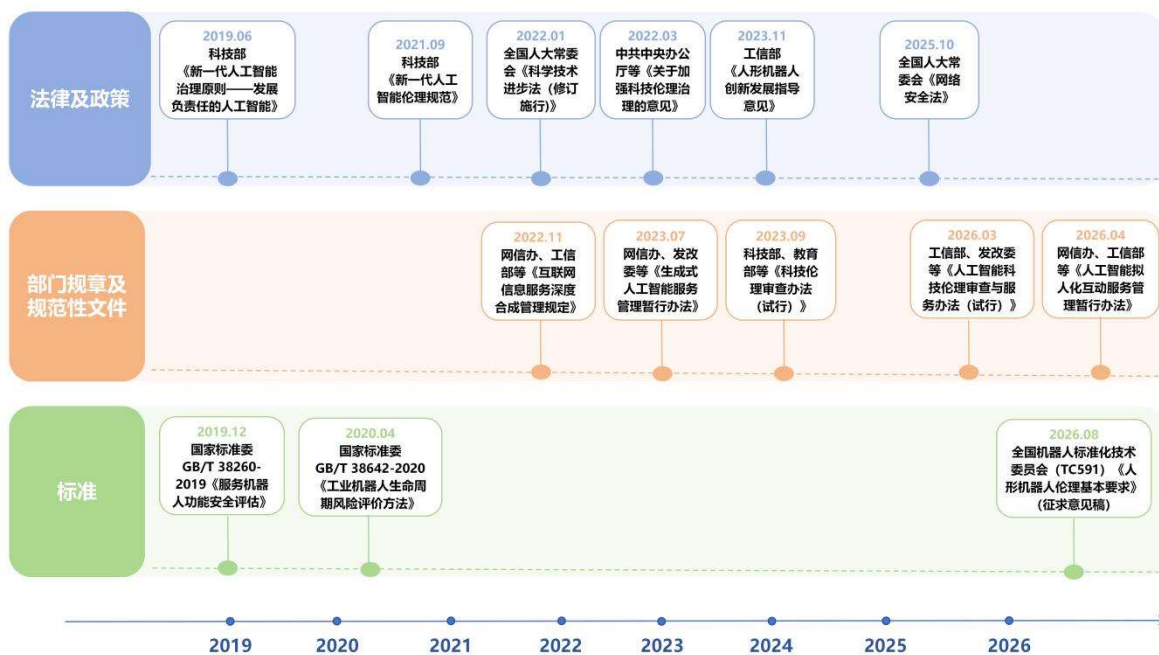


图 2-1 国内涉人工智能与机器人领域科技伦理重要文件时间分布图

总体来看，中国已经构建起覆盖价值原则、专项应用服务治理、科技伦理审查、产业政策与标准预研的人形机器人治理框架，为治理工作提供了重要的基础，但全生命周期伦理评价体系的构建，以及人形机器人专项合规要求的整合仍在推进过程中。

2.2 国际组织及标准机构

2.2.1 国际组织提供共同价值与政策语言

2021 年 11 月，联合国教科文组织（UNESCO）发布《人工智能伦理问题建议书》（Recommendation on the Ethics of Artificial Intelligence），获得所有成员国的一致通过。作为全球首个具有普遍覆盖性的 AI 伦理规范文本，该建议书明确以人权、人的尊严、环境与生态可持续性为根本价值根基，并将比例性与不伤害、安全保障、公平与非歧视、可持续性、隐私保护、透明性、问责制以及公众参与等核心伦理原则，系统性地转化为数据治理、教育、健康、文化、传播与环境等具体政策领域的要求，为不同法域提供了相对稳定的价值话语体系。

2019 年 3 月，电气与电子工程师协会（IEEE）正式发布《人工智能设计的伦理准则》（Ethically Aligned Design），将人权、人类福祉、数据治理和责任等伦理考量融入自主与智能系统的设计之中。同年 4 月的欧盟高级别人工智能专家组《可信人工智能伦理指南》，将合法、合乎伦理与技术稳健列为可信人工智能的三个原则，并提出七项要求，体现了专业共同体将伦理考量前移至研发环节的努力，既有助于研究者和工程人员识别设计选择的价值影响，也可为企业原则制定和标准研制提供参考。

国际原则为人形机器人科技伦理治理提供了基础性的价值指引，当前国际组织已经解决了价值话语分散的问题，但尚未独立完成面向人形机器人的工程化与制度化转译工作。人的尊严、安全、公平、隐私与自主等核心价值不会因为人工智能系统拥有机械身体就发生变化。人形机器人的特殊性体现在上述价值受影响的具体路径与方式中。如透明性原则需要进一步细化为对人形机器人的非人身份、能力边界与社会关系角色的具体说明，而安全原则需要转化为对人形机器人感知、决策与执行全流程连续运行的控制要求。

2.2.2 标准机构建立管理、风险与影响评估框架

2017 年，ISO/IEC JTC 1 决议设立 SC 42 人工智能分委会，主要负责开展人工智能管理体系、数据治理、系统测试及可信赖性等方面的国际标准化工作。2023 年，SC 42 发布 ISO/IEC 42001《人工智能管理体系》（Artificial Intelligence Management System）与 ISO/IEC 23894《人工智能风险管理指南》（Artificial intelligence Guidance on risk management），其中，ISO/IEC 42001《人工智能管理体系》聚焦于为组织建立、实施、维护和持续改进 AI 管理体系提供系统化框架；ISO/IEC 23894《人工智能风险管理指南》则为 AI 系统全生命周期中的风险识别、分析、评价、处置与监控提供方法论指导。2025 年 5 月，SC 42 进一步发布 ISO/IEC 42005《人工智能系统社会与人类影响评估》（AI system impact assessment），帮助各组织识别、评估和管理人工智能在其整个生命周期

中对人类、社会和组织的影响，提供影响评估的原则、流程与方法框架，涵盖人权保障、社会公平、文化多样性、就业结构变迁及弱势群体保护等维度，并将评估嵌入系统设计、开发、部署、运营和退场全过程。

机器人标准化工作过去长期聚焦在工业和服务机器人的安全、性能与测试方面。比如欧盟机械法规将协调标准和合格评定作为产品准入制度的重要实施环节；国内 GB/T 38642-2020 和 GB/T 38260-2019 分别为工业人形机器人生命周期风险评估和服务机器人功能安全评估提供了基础性标准。

上述标准既能够为人形机器人搭建统一的管理框架，也便于各企业去复用自身已有的质量、信息安全和风险管理能力。但这类标准的跨行业通用性也决定了它的适用边界，标准仅提供通用方法框架，但是产品风险能力画像、风险形成机制、风险评估尺度和测试验证的证据清单等，仍然需要围绕人形机器人的专用场景进行具体补充。即便企业采用上述标准管理体系也只能证明其建立了相应合规流程，不能直接证明人形机器人在具体场景中的伦理合格性。

2.3 现有治理框架对物理行动的覆盖情况

2.3.1 人工智能治理规则：覆盖决策环节但执行环节不足

当前人工智能治理规则对决策环节的覆盖已较为完整，能够有效处理模型数据质量、性能边界及失效模式等相关问题，但总体而言现有的规则体系仍是以数字输出为核心规制对象，在应对具身智能时存在三方面缺陷。

第一，人类监督要求可能需要在具身条件下重新定义，单纯依靠传统的界面确认方式不足以构成有效的人类监督，现场人员必须在机器人动作发生之前获取充分的信息，并且具备实际的中止能力，还预留充足反应时间，三者缺一不可。

第二，现有准确性与稳健性的评价要求缺少后果强度维度，相同错误率在纯文本输出场景与具身场景中造成的后果差异十分显著，模型准确率高并不等同于应用在物理动作执行场景时会安全。

第三，以新加坡智能体 AI 治理框架为代表的现有智能体监管框架，虽然已经在具身智能场景时也多数能够复用，但是存在操作可枚举、权限可分级、执行可撤销的适用前提，在物理动作执行场景中并不完全成立，原因在于数字智能体的权限管理机制无法直接适用到具身智能体的物理执行环节。

2.3.2 数据保护规则：覆盖数据处理过程但不覆盖处理后果

现有的数据保护规则对人形机器人的数据处理活动构成直接约束，包括合法性基础、目的限制、最小必要等核心要求。但现有规制范围仅限于数据处理活动，并不覆

盖机器人利用个人信息形成物理动作所引发的后果，尽管感知数据的采集质量会直接影响动作安全。数字世界的数据处理是一个线性的过程：采集、传输、存储、加工、使用、共享、删除，数据本身并未改变，只是数据的流向在发生变化，治理的对象可以落到具体的数据字段和 API 接口请求上。但具身智能的数据处理本质上变成了感知、推断、决策、执行、记忆构成的“信息的输入和输出”循环，它处理的数据范围扩展至整个物理环境空间，以及环境里的所有人，传感器持续感知周围的一切，从采集的数据中推断出结论（比如对方是主人还是一个访客），作出决策并驱动物理行动，行动改变环境之后又被重新感知。

所以个人信息不再只是在数字世界里流转，还会改变机器如何进入物理空间、接近人员和执行动作。数据治理的对象也必须从数据字段和接口请求，扩展为感知、推断、行动、记忆的所有环节。数据保护规则与物理安全规则在制度层面可能构成两套相互独立的保护规则体系。当前部分企业可能会将完成个人信息保护影响评估视为风险评估已经完成，但忽视了物理层面的安全性审查。

表 2-1 现有规则对人形机器人不同环节的覆盖情况

环节	主要覆盖工具	覆盖程度	边界
感知	机械与机器人安全标准（传感器性能）、数据保护规则（采集范围）	部分	传感器性能和数据采集分属两套规则；超出标定环境时如何处理，没有统一要求
决策	人工智能治理规则、风险管理标准	较完整	只看数字输出，不区分后果严重程度；模型准确率高，不代表具身部署安全
执行	机械与机器人安全标准	薄弱	模型判断出错和硬件故障的处理方式不一样；智能体的权限管理覆盖不到物理动作
人的反应	机械安全标准（作为假设条件）	薄弱	标准假设用户受过培训，但家庭等场景中不成立；拟人外观如何影响安全距离，未被纳入评估
风险传导与证据对齐	无专门工具	缺失	每个部件都处理合格不代表整体安全；跨部件的联动验证没有统一标准；各自产生证据，缺乏统一的编号、时间基准和保存规则，无法互相对应

2.4 现有治理框架对拟人交互的覆盖情况

拟人化引发的关系风险是随着时间推移逐步形成的，通常可划分为首次接触、重复互动、关系形成、关系变化与终止四个阶段。与物理行动风险大多在某一特定时点产生的特点不同，关系性风险一般是在长期互动中缓慢积累形成，且每一阶段的表现往往都可能符合该阶段的合规要求。基于此，下文将从现有治理工具是否覆盖上述四个阶段、风险判断依据是基于单次行为还是累积行为状态、用户是否拥有可行的关系退出路径等多个维度，系统分析当前人机交互风险的治理缺口。

2.4.1 身份透明：覆盖首次接触但对累积影响有限

身份透明是现有治理框架中覆盖最明确的要求（如欧盟《人工智能法案》第五十条、中国《人工智能拟人化互动服务管理暂行办法》等），直接回应了用户首次接触AI服务阶段的身份混淆风险。但这一机制的作用存在三项明显限制。

其一，当前身份告知的时点与频率通常围绕单次交互设计，首次告知虽能满足初始合规要求，却无法保证用户在数月的持续互动中，始终对AI的非人身份保持清晰认知；

其二，身份披露不等同于能力边界披露，即使用户知晓交互对象是AI，也无法仅凭这一点准确判断AI的理解深度、专业水准以及适用范围；

其三，身份告知和产品营销宣传存在一定冲突。产品宣传往往倾向于把AI机器人塑造为“陪伴者”“伙伴”这类拟人化形象用户购买的预期中可能本身就自带一部分“希望混淆”的动机，而产品服务条款则会明确强调其非真人属性，现行规则主要对服务条款内容作出约束，并未对营销表达与服务条款之间的一致性作出直接规范，这可能会削弱身份透明机制的实际效力。

2.4.2 操纵性禁止：覆盖显著情形但偏向单次行为

欧盟《人工智能法案》第五条明确禁止操纵性技术及利用特定群体脆弱性的行为，中国《人工智能拟人化互动服务管理暂行办法》也对情感引导和未成年人保护设置了专门要求，为拟人交互划定了实质性边界。但上述禁止性规定的适用，通常需要以识别和认定操纵性技术、行为扭曲及重大损害可能性为前提，这类规则对于单次或短期内可识别的操纵行为约束效果较好，针对持续迎合用户观点、逐步扩大排他性表达这类长期缓慢积累的影响却难以适用。此外，现有弱势群体保护条款的覆盖范围还需进一步明确，人形机器人应用场景中，用户脆弱性的来源更为多样，组织从属关系中的服务对象、独居老人、认知功能退化者等群体是否落入现有规定的列举范围，仍需要结合具体个案判断，目前尚缺乏统一明确的认定标准。

2.4.3 数据保护：覆盖记忆处理但不覆盖关系本身

数据保护规则对关系性风险具备部分覆盖能力，例如记忆构建、用户画像生成与个性化配置，均是当前较为常见的个人信息处理活动，用户的删除请求事实上能够中断关系记忆。但数据保护规则的规制对象是数据而非关系本身，用户与平台形成的使用依赖、对产品角色的熟悉感、在双向反馈中养成的表达习惯等都不会因个人信息被删除而消失。反过来，当用户希望保留原有关系仅更换设备时，可携带权仅能处理数据格式转换与传输问题，无法保障迁移后角色表现的连续性。这种规制错位导致现有规则仅能处理关系的“数据载体”问题，无法规范关系本身的形成、变化与终止，因此本白皮书提出需要区分退出、删除、重置、迁移四项机制（详见第六章 6.1.5）。

2.4.4 未成年人与特定场景：覆盖特定主体但场景外延不足

现有未成年人保护规范以及针对学校、医疗、养老等特定应用场景的制定的专门法规已经对拟人化交互中的相关风险作出了规制，这类规则通常按照主体或场景界定适用范围，边界相对清晰。但在实际运行过程中，规则的具体适用范围仍然需要依托个案进行判断。例如，面向普通消费者开发的产品可能会被未成年人使用，家庭场景中的机器人需要同时服务成人与儿童，开放环境中的交互对象也无法在产品设计时提前预设。此外，当行业管理既有的专业责任要求与产品端的 AI 治理规则同时适用时，目前缺乏有效的衔接机制。这一问题在人工复核环节尤为突出，人工干预的节点、执行主体、记录所依据的标准都没有明确规定，最终可能导致两类责任规范在具体操作层面难以落地。

2.4.5 第三方主体：保护不足

第三方主体指未主动选择与机器人互动，但因处于机器人的感知或行动范围内而被动卷入实质性交互的个体，如商场中的顾客、家庭中的非主要使用者、养老院中的其他老人。现有规则对第三方主体保护不足，其根源一定程度上在于制度设计层面的缺陷。传统的透明度管理普遍依赖告知机制，但第三方主体往往无法及时获得告知，或者难以注意到相关提示信息；传统的同意与拒绝机制依托用户账号体系运行，但第三方主体通常并不拥有对应的账号；机器人的无感采集和便利交互特点，容易使第三方主体并不知晓自身数据已经在被采集。当前针对第三方主体规范工具仅有公共场所摄像提示和产品安全警示两类，均未触及人机互动方面的影响。此外，第三方主体中的弱势主体包括但不限于儿童、行动受限者以及认知受限者，面临的保护缺失问题可能更为突出，目前相关保护工作主要依靠部署者或运营者自行把握保护尺度，尚未形成统一可操作的标准，也缺乏制度层面的刚性约束。

综合来看，关系性风险的治理难点在于关系的持续性。现有规则能够回应“用户

是否知晓对方是机器”问题，但这只是关系性风险的治理起点，难以确保几个月之后用户的这一认知仍然有效。同时，监测关系性风险需要长期的数据支撑，但持续推断用户心理状态又可能引发新的隐私与自主风险，这一矛盾需要在风险评估方法层面进行审慎的处理。本白皮书并不主张因为存在风险就否定拟人化设计，拟人语言、动作、表情有助于降低交互门槛，稳定的人格设定也有助于建立服务连续性，治理的关键并不在于产品拟人化程度的高低，而在于区分两类设计，一类是满足服务功能所需的拟人表达，另一类是利用用户的脆弱处境放大依赖的设计，区分的具体判断依据可以是关系角色、持续时长、用户理解能力和退出条件等综合来定。

表 2-2 现有规则对人形机器人与人交互不同阶段的覆盖情况

阶段	主要覆盖工具	覆盖程度	边界
首次接触	AI Act 第 50 条、拟人化互动办法（身份提示）	较完整	身份披露不代表能力边界披露；用户预期、营销话术与条款内容之间实际缺乏一致性
重复互动	AI Act 第 5 条（操纵与脆弱性利用禁止）	部分	判断标准侧重于单次、可识别的具体行为，对于缓慢累积形成的影响则难以适用
关系形成	数据保护规则（关系记忆的采集与删除）	薄弱	现有规则只覆盖关系的数据载体本身，并未覆盖用户对机器人产生的依赖、关系被替代、决策与行为让渡等问题
关系变化与终止	数据保护规则（删除、可携带）	薄弱	针对人格变化、迁移后的连续性、停服过渡等情形，尚无专门规则；数据可携带权无法覆盖机器人的关系处理需求
累积状态的判断方法	无专门工具	缺失	缺乏以时间为维度的评价方法；而持续监测用户心理状态又会带来新的隐私与自主风险
第三方的关系影响	无专门工具	缺失	告知与同意机制以用户拥有账号为前提，无账号用户难以被纳入治理流程

2.5 现有治理框架不足的总结

综合本章对全球科技伦理治理各类工具的梳理总结，以及现有框架对人形机器人

特有伦理风险覆盖情况的分析，我们认为，现有框架应用于人形机器人科技伦理治理时，主要存在以下四类不足。

一是制度识别层面的对象识别不足。人形机器人作为一个新兴技术产物，在现行的治理框架中被拆解成了多个碎片化的规制对象，不同的制度规则仅能覆盖人形机器人的不同方面。具体而言，数据保护规则只是将其视为一连串的个人信息举报活动，人工智能规则只是将其视为一个算法模型或系统，机械产品法规只是将其视为机械设备，网络安全法规则只是将其视为一个联网产品，消费者保护和产品责任制度则侧重于识别和判断是否构成交易关系与损害后果。这些风险识别方式和规制路径在各自领域内虽然都具备合理性但彼此割裂，未能将人形机器人作为一项完整的、具有社会属性的新兴技术产品加以整体规制。

二是方法论层面的风险评估机制不足。现有的治理框架虽然在各自适用范围内具备一定完整性，但缺乏一套贯穿数字判断、物理执行和人机交互反应全链条的风险评估方法与评价证据链。首先是各类规则的适用前提存在差异，机械安全法规关注可预见的危险和产品状态，数据保护规则关注处理目的与对个人权益的影响，人工智能治理关注算法特性和决策逻辑。但是风险恰恰在这些治理框架的边界之间传导，现行的任何单一规则难以完整记录并追踪这一跨环节的风险传导过程。其次是从数字判断到物理执行缺乏统一评估。即便人形机器人的零部件或软件子系统分别通过单项合格认定，仍可能在特定时序、环境和权限配置的组合下产生不可接受的后果。如果人类监督仅停留在界面确认按钮或机器人本体的急停按钮层面，现场操作人员往往未必拥有足够的信息和反应时间来有效处置突发状况。最后是多主体的伦理责任分配可能因为缺乏可追溯性和完整性而难以确定。第三方模型提供商、硬件供应商、整机厂商、部署者、运营商和用户可能分别控制风险形成过程的不同环节。如果事件记录无法还原事发时的模型版本、系统配置、感知数据、指令序列和现场环境条件，责任认定往往只能停留在主观判断层面，难以客观评估各方对风险的控制力和影响程度。

三是时间维度的关系治理不足。现有规则以单次互动和相对比较清晰的显式后果为判断依据，对累积影响、关系演变、关系终止及由此产生的隐性后果覆盖不足。其一，用户对人形机器人的信任是在多次互动中逐步形成的，如果机器人能够记住用户或家庭成员之间的共同经历、保持稳定角色，甚至主动与用户互动，用户往往会在行为上逐步按照现实的人际关系模式来回应。其二，关系的变化与退出缺乏明确的制度安排。模型的升级可能改变机器人的角色表现，跨设备迁移可能丢失或改变人机共同记忆，企业停服则会同时中断产品功能与人机关系。另外，用户能否退出、删除、重置或迁移人机关系，服务商是否需要提供过渡性安排，以及护理、教育和陪伴等角色中哪些人类责任不能由产品替代等等，这些问题都需要超越现有制度安排。其三，关系影响存在证据滞后性，且关于用户的数据监测本身存在一定限制。物理碰撞可以在

事件发生后即时记录，但是依赖形成、关系替代和决策让渡往往是一个缓慢积累的过程。如果为了监测这些影响而持续推断用户的心理状态又可能引发新的隐私与自主风险，在这一点上如何平衡，当前的治理框架尚未形成成熟的方案。

四是动态适配不足。这一问题的深层原因在于，一方面，技术演进与产品迭代始终处于动态变化之中，相关标准和政策出台往往滞后于市场。在人形机器人相关标准尚处于研制阶段时，企业往往需要从通用人工智能、机械安全、数据治理和科技伦理等各类文件中自行拼凑合规要求，导致不同企业的实践路径差异较大，难以形成可比较、可统一的操作规范。另一方面，机器人交付使用后仍然可能持续经历算法模型更换、软件版本迭代、权限范围扩展以及应用场景迁移等变化，拟人化角色设定和用户关系形态也可能随运营策略动态调整。上市时经认定合格的产品版本和功能配置，并不足以确保其未来状态持续合规。尽管现有制度规则已开始引入上市后监测、事件报告和持续改进等要求，但治理供给仍显不足。如何在动态发展中保持治理的适应性与前瞻性仍是一个亟待回应的制度命题。

上述四类不足相互关联，分别指出了当前治理框架中的风险识别、评估方法以及风险随时间的动态演化等不同维度的不足。人形机器人需要构建一套清晰的伦理风险治理框架。

表 2-3 人形机器人治理框架不足汇总

适用性不足	制度层面的问题
对象识别不足	产品被分别识别为人工智能、机械、数据处理活动或服务，缺乏对物理行动、拟人表达和持续关系的整体刻画
评价机制不足	分域规则缺乏贯通数字判断、物理执行、人的反应与伦理影响的评价体系和证据链
关系治理不足	透明、隐私和监督规则仅覆盖部分单次互动，对长期信任、依赖、关系替代、角色变化和服务终止的安排不足
动态适配不足	标准、法律和实施能力不同步，静态结论难以覆盖版本、场景和主体变化

第三章 人形机器人伦理风险成因

事实上，人形机器人伦理风险很大程度上来自于机器人所具备的技术能力和部署条件。技术本身并不一定会直接带来风险性后果，只有当这些技术能力和部署条件能够改变权利、控制、责任或社会规范时才构成了伦理风险的来源。

3.1 人形机器人伦理风险的技术来源

人形机器人集机械执行、网络通信、人工智能和拟人交互于一体，可以兼具物理行动、数据处理、人机关系等多重功能属性，在物理世界它能够移动、抓取并与人类直接接触，在数字世界它能够接收、采集并更新数据，在与人类互动时它可以识别、生成并动态调整行为以适应当前情境，也可能在长期使用中记忆用户特征进而稳定地扮演特定角色。这三重能力的深度融合，在拓展机器人功能边界的同时也催生了深层次的伦理隐患。为了厘清伦理风险的根本来源，首先需要了解人形机器人所依托的技术以及技术风险的传导路径。本白皮书为方便理解，将人形机器人的技术风险按照五个技术能力层面逐层展开，每个层面代表一类主要的技术维度。需要说明的是，这样划分仅仅为了直观显示不同技术维度下特有的风险形态，以及不同层面的技术之间的风险叠加、放大、衍生、质变等风险传导路径，为后续分析人形机器人的伦理风险形成机制奠定基础，并不代表各个层级之间存在明显的上下层级关系（虽然从技术成熟和规模应用的时间顺序上有一定先后次序）。

第一层是机械能力层。代表传统以机械能力为主的工具性实体。人形机器人的机械结构与传统机电设备类似，由关节、手臂、腿部、手部等执行机构、电气系统、结构材料及底层控制程序构成，主要按照预设规则或外部指令完成指定动作。即便配备了触觉或距离传感器，且能够在触发条件时停机或局部避障，只要不能根据整体环境自主选择任务目标和行动策略，其核心能力就仍停留在机械层。该层面的风险主要源自零部件失效、过热、漏电、误启动、制动失灵或机械臂异常运动，这些问题可能导致功能失常并造成物理损害。

第二层是网络能力层。代表具备网络连接和数据处理能力的机械实体。人形机器人接入通信网络和云端服务后就不再是孤立设备，而成为网络中的一个数据处理节点。环境图像、语音、位置和操作记录等数据可实时采集并上传，软件和控制程序也支持远程更新。与此同时，风险来源也随之扩展至通信网络、云服务接口、第三方组件和远程控制权限，攻击者无需接近机器人本体即可利用漏洞窃取数据、篡改指令甚至接管系统，风险类型由此从传统的物理安全延伸至网络与数据安全。风险来源、风险类型的扩展与演变，又进一步从三个维度放大了风险的影响范围。空间上，风险从机器

人设备本身蔓延到云平台、移动端和第三方服务；时间上，出厂安全状态不能保证覆盖人形机器人的整个生命周期，软件更新、配置变更和组件停止维护都可能引入新的风险；规模上，一次远程攻击可能同时波及分布在各地的多台机器人。

第三层是智能能力层。代表具备网络连接和数据处理能力，还有一定的智能计算能力，但不一定具备机械结构，不具备具身行动能力的产品或服务。人工智能大模型系统借助数据参与识别、预测、分类、推荐等决策环节，由此带来的风险已不止于输出错误，还涉及数据偏差、群体歧视、算法黑箱、自动化偏见、虚假信息，乃至对人的行为操纵。在这一层面，数据风险被算法机制显著放大。传统数据安全问题如信息泄露的损害相对明确、可追溯，但当带有偏差的数据进入自动化决策系统，偏差就不再是一次性的，而是会通过模型训练、反馈数据和决策流程被持续强化，歧视性结果的影响范围会不断被放大，风险也随之积累。

第四层是具身能力层。代表具备人工智能的感知、推理与决策能力的可运动、可操作的实体。系统通过摄像头、雷达、触觉传感器等感知环境，再借助机械臂、轮式底盘或双足结构作用于现实世界，由此从决策参与者变成了物理执行者。数字世界的人工智能系统错误通常停留在信息层面，识别一张虚拟图像出错，输出的只是一个错误结果，影响可逆、范围可控；但具身智能系统的错误直接作用于物理世界，感知一旦失误，就可能转化为碰撞、挤压、跌落等即时且不可逆的物理伤害。

第五层是拟人能力层。在具身智能基础上，人形机器人进一步融入了外形、表情、语言、视线、情绪与角色设定，使机器从物理执行者演变为具有社交属性的交互对象。它可能以家庭成员、照护者、教师、同事、朋友或情感伴侣等身份参与人类日常生活，成为一个能长期记忆个人信息、模拟关怀、回应情绪并介入社会关系的技术实体。由此，拟人化让人形机器人的风险从产品风险升级为新型的人机关系风险。当人机关系从工具性使用转向社交性互动，其影响也相应地从物理安全、数据权益和行为控制延伸至身体伦理、情感认知、社会角色认同乃至自我认知等更深层领域。



图 3-1 技术风险演进示意图

以上五层技术能力描述了人形机器人的主要技术能力组合，不同的技术能力层在机器人设计时可能是一个叠加或者多种组合的方式。例如工业人形机器人往往物理执行能力很强，但只接受预设指令；轮式服务机器人没有完整的拟人结构，但能联网、记忆用户并进行多轮对话。此外，随着软件更新、模型接入和应用扩展，同一人形机器人产品的能力组合也会不断发生改变。能力组合的不同，决定了风险的形成与演变路径。本白皮书针对以上五个技术能力层归纳了三种具有代表性的风险形成与演变路径，三者的形成机制性质不同，需要不同的风险治理策略。

第一种风险形态是底层（如机械层）基础风险持续存在并叠加。机械层风险不会因为人形机器人有了智能层等上层能力就自动消失。联网可能扩大风险的影响范围，人工智能可能改变机器人的行为逻辑，具身动作可能把错误变成物理伤害，拟人交互又可能左右人们对这些伤害的认知。风险在每一个能力层都有特殊的形态，传统的合规判断无法跨层套用，必须逐层评估、逐层论证。

第二种风险形态是误差跨层传导衍生新的风险。前述五层模型中，通常情况下上一层的输出可以构成下一层的输入，比如传感器误差进入模型判断，便成为人形机器人判断错误的前提，判断误差传递至执行环节便转化为执行偏差，而执行偏差又可能进一步干扰现场人员的判断与反应。此类风险传导方式可能表现未，每一个环节的误差单独来评估和审查都是在容差范围内（比如传感器误差未超限值，模型偏差未逾阈值，通信延迟与机械执行误差各自处于合理区间），但这些微小误差一旦在特定时序与空间条件下叠加汇聚，便可能突破系统整体安全边界，使人形机器人在物理层面出现越界行为。

第三种风险形态是纠错能力被削弱导致风险被放大。区别于前两种风险形态，此类风险形态并非直接提高风险发生的概率，而是表现为降低风险被发现、被纠正的可

能性。比如，拟人层能力使得人形机器人与用户建立信任，使用户在日常交互中逐渐放松对机器人行为的主动核查；智能层通过机器人流畅、肯定的输出，使得用户倾向于减少质疑；网络层则通过远程更新机制，使机器人的功能与行为逻辑变化对用户基本不可见。三者共同作用之下，技术缺陷得以在隐蔽中持续存在，发现越晚，纠正越难，后果也就越严重。

对于机械风险、网络风险、人工智能风险，现行治理体系已形成相对成熟的制度框架，但其运行边界仍停留在信息处理、数据加工与决策输出阶段，尚未在开放环境中形成持续、自主行动所需的完整闭环。具备这些技术能力的物理设备所产生的伦理风险大多数仍通过产品安全、网络安全、数据保护和算法治理等规则来识别与应对，但尚未触及人形机器人因具身性、拟人性和关系性而产生的深层伦理风险，这类风险恰恰是人形机器人区别于普通人工智能产品的独特所在，也是当前科技伦理治理亟待回应的核心命题。

3.2 人形机器人的特有伦理风险

人形机器人特有的伦理风险主要沿现实世界中物理行动、社会认知和持续关系三条路径形成。具身性使智能系统能够进入现实空间并改变身体、物品、设备和环境状态；拟人性使机器人能够再现人的身体、身份、情感、意图和社会角色，从而改变人对技术的理解；关系性使多次互动形成共同历史、稳定人格、角色期待和退出成本，使影响在时间中累积。三类风险属性，既可独立产生伦理风险也可能有多种组合形式。关系性风险没有在 3.1 节的技术能力层单独体现，它是记忆、主动性、人格连续和角色定位等设计因素，在时间积累过程中逐步形成的隐性风险。

3.2.1 具身性

具身智能将人工智能的感知、推理和决策能力赋予可运动、可操作的实体，使人工智能取得现实行动能力，能够直接作用于物理世界。数字层面的错误输出、数据偏差与网络攻击，因此有可能转化为物理层面的伤害。机器人还可能取得进入空间、触碰身体、处置物品、控制设备和执行制度规则的权限。碰撞、夹挤、跌倒、零部件故障首先属于产品安全问题，伦理分析关注的是其背后的授权、控制与责任分配。

（一）人形机器人常见具身能力

表 3-1 人形机器人常见具身能力

具身能力	说明
具身感知	持续感知人员、身体、物品、空间和环境状态，并将现实空间转化为数字信息

具身能力	说明
具身移动	接近、跟随、避让、占据或阻挡空间，改变人与机器、人与物之间的相对位置
具身操作	触碰、搀扶、搬运或约束人体，抓取、递送、处置物品，控制门锁、家电和其他设备
具身介入	以照护、教育、安保、管理或现实代理的方式进入家庭、组织及公共流程
具身表达与示范	通过现实动作展示舞蹈、格斗、亲密、服从、支配、惩罚等可被观看和模仿的行为示范

（二）具身性伦理风险

具身性伦理风险主要关注谁有权决定、谁能拒绝和中止、谁承担适应成本、哪些群体得到的保护不足，以及产品设计与控制权、受益度与责任是否匹配。具身性伦理风险来源于人形机器人的具身能力，具身能力给人形机器人带来的伦理风险主要体现在以下几个方面：

表 3-2 典型具身性伦理风险

序号	直接风险	典型表现
1	身体自主与接触边界	即使安全的动作也可能缺乏正当授权。机器人触碰、护理、康复或亲密接触时，应区分初始授权、当下同意、撤回、代理同意与非语言抗拒
2	行动自主与实质控制	自动执行可能把人的选择限缩为接受既成结果。形式上的确认、急停或人在回路中不代表人具有足够信息、反应时间和真实接管能力
3	空间自主与不可回避暴露	移动、跟随、占位及持续感知可能迫使人避让、改变行为或失去私人空间。家庭成员、访客、路人等未必能关闭或回避机器人
4	现实权限越界与目的漂移	清洁、收件、提醒等低风险权限可能扩展到卧室、开门、药物、支付和物品处置；软件更新和权限组合还会改变原授权含义。财产权益与家庭秩序是本类风险的重要场景
5	差别影响与风险适应转嫁	标准的人体、标准的动作和标准的环境等不足的训练样本或假设，可能让儿童、老人、残障者及其他行为者承担更高风险；部署者还可能把理解、避让和适应机器人的责任转嫁给现场人员
6	重大行动及照护	即使系统准确执行，被交付给机器人的任务本身也可能包含不宜完

序号	直接风险	典型表现
	责任不当委托	全技术化的价值判断，如约束、看护、安保、危机处置及涉及基本权益的不可逆行动
7	具身表演与行为规范塑造	格斗、羞辱、性化、服从和支配动作可被娱乐化、观看和模仿，可能影响公众对暴力、身体和社会角色的理解；相关长期行为迁移证据仍有限
8	责任分配与救济失衡	模型提供商、制造商、部署者、使用者和现场人员共同完成行动，责任却可能集中到缺乏信息和控制权的近场人员；第三方还可能找不到投诉、暂停和救济入口

（三）典型具身性风险形成机制

表 3-3 典型具身性风险形成机制

序号	风险形成机制	示例
1	持续感知→共享空间数据化→第三方进入→聚合画像→私人空间期待改变	家用机器人每天在家庭活动并将日常采集的数据传输至第三方处理或存储，第三方可以推断房屋面积、作息规律、家庭成员数量等用户画像，向外部提供生活习惯数据
2	接触能力→人的身体成为操作对象→概括同意被长期化→当下撤回或抗拒被忽视	护理人形机器人根据需求设定或家属同意每天定时帮老人递水，老人拒绝被服务，机器人短暂停顿后仍按既定程序继续提供服务
3	低风险授权→成功累积信任→拓展权限组合或默认续权→提醒扩展为执行	机器人每天提醒老人吃药，开灯、播报天气，了解用户习惯后能够直接把药盒摆到桌上，并自动记忆了用户存在辅助喂药需求，但老人和家属可能从来没有明示同意过该项服务
4	安全或健康目标→系统判断用户受益度→提醒升级为限制和替代决定	机器人依据用户或者其监护人设定的模糊安全与健康目标，起初只是提醒用户吃药、休息等。假如提醒不起作用，可能升级为劝阻或者进行物理限制，包括通知其监护人，导致用户反感
5	标准化算法模型→群体适配不足→风险与服务质量差异	商场导览人形机器人在测试中避让成功率达 99%，但对轮椅使用者和身高一米以下的儿童两类人的训练数

序号	风险形成机制	示例
	→实质性不平等	据不足，导致避让不成功
6	行动意图不透明→人必须预测和避让→第三方主体承担额外成本	写字楼中的配送机器人转弯时从不示意，快递员、保洁员和从它身边经过的员工等第三方主体不得不放慢脚步观察，并考虑适时躲避或主动给机器人让路
7	自动任务扩大→人类监督减少→价值判断不能对齐→事故后责任不清	仓库安防机器人遵循夜间禁止人员进入的安保职责，并顺畅地执行了很长时间，某晚它拦下了一位急需给值班人员进行医疗救治的陌生医护人员
8	现实动作展示→观看与模仿→形成暴力、性化或服从惯性认知	某游乐场让儿童观众上台与格斗机器人进行对打的娱乐游戏，一段时间后幼儿园老师发现班里孩子开始模仿那些对打动作，其中一些孩子开始用同样的方式推搡同学，并声称“反正它又不会疼”

3.2.2 拟人性

拟人性使机器获得社会表征与认知影响力。拟人性不限于人脸和人形外观，声音、语言、目光、表情、动作、人设、专业角色以及体温、触感和呼吸等拟生命特征都可能使人按照理解真人的方式理解机器人。拟人化本身不是风险，它可以降低交互门槛并改善无障碍体验。真正的伦理问题出现在拟人呈现与系统真实身份、能力、情感和责任之间产生差距时，或者拟人身体及身份表征直接贬损、物化或复制现实中的人和群体。

（一）人形机器人常见拟人能力

表 3-4 人形机器人常见拟人能力

拟人能力	说明
身体再现	通过面容、皮肤、毛发、身高、体型、性征、四肢及身体细节再现人的身体
身份模拟	通过肖像、声音、动作、服装、标识和资料模拟特定真人、机构或社会身份
社会表达	通过语言、目光、表情、姿态、礼貌、示弱、权威及情绪表达触发社会反应
心智表征	通过第一人称感受、意图、承诺、价值和自主决定表达呈现似有心智的主体
角色承载	持续或阶段性扮演医生、教师、护理者、朋友、伴侣、英烈或历史人物等角色
拟生命呈现	通过体温、触感、呼吸、心跳、眼泪和触摸反馈模拟生命体征和身体反应

（二）拟人性伦理风险

人们往往会根据表情、语言、动作和回应方式来推断交互对象的内在意图，当人形机器人保持目光接触、用自然语气回应，或者在用户低落时及时表达关切，用户就会下意识地调用理解真人行为的认知来解读这些信号。机器是否真的有情感和用户是否在人机交互中产生了真实的情感反应是两个不同问题。

表 3-5 典型拟人性伦理风险

序号	直接风险	典型表现
1	人机身份与信息来源误认	用户无法判断交互对象是否为 AI、是否有真人实时参与、输出代表谁，或者误把合成形象、声音和内容当作本人真实表达
2	能力、资格与社会权威误认	流畅语言、制服、称谓和专业姿态可能使人高估理解、记忆、保密、诊断、报警、评价和命令权限
3	情感、意图、意识与人格误认	用户把模拟关心、痛苦、欲望、承诺和选择理解为真实体验、稳定人格或独立道德主体，并据此调整行为
4	社会性影响与自主决定受损	寒暄、同理、示弱、权威、群体压力和情绪诉求可能改变信任、服从、披露、购买和反馈，但拟人化并不必然导致过度信任或敏感信息披露
5	身体、群体与社会角色的不当表征	涵盖性别化、性化与物化，种族、年龄、残障及职业刻板印象，以及将年轻、纤细、健全、无瑕等单一身体定义为优越或正常
6	特定真人及人格身份的非授权复制	未经适当授权复制在世者、伴侣、儿童、公众人物、逝者、英烈或历史人物的面容、声音、身体和人格资料；AI 标识不能替代人格授权与公共利益审查
7	拟生命真实性、感知不适与场景化惊吓	体温、触感、呼吸等可能造成生命和感受误认；虽然高度仿生但与真人的眼部、声音、动作不一致可能引发怪异感；夜间、卧室、病房等场景还可能造成不可回避的威胁性误认和惊吓

（三）典型拟人性风险形成机制

拟人性风险有三条典型的风险形成路径，一种是表征性路径，即拟人身体和身份如何再现性别、族群、年龄、身体、特定真人、逝者和社会角色，并直接影响尊严、平等和人格利益；一种是归因性路径，即用户如何把拟人表达解释为身份、能力、情

感、意图、意识和责任，并进一步改变信任、披露、服从和归责；还有一种是感知性路径，当机器人表现出高度拟真、跨模态不一致、突现和不可回避暴露直接引起怪异、惊吓、厌恶、警觉和空间回避，此类恐怖谷效应带来的不适是对拟人性实体的警觉反应。

表 3-6 典型拟人性风险形成机制

路径	风险形成机制	示例
表征性 路径	第二性征与性化身体设计→身体被呈现为可观看、购买、服从和支配的对象→现实性别关系与身体尊严被贬损	仿真伴侣机器人以女性、男性身体突出的第二性征为卖点，营销中突出永远顺从、不会拒绝或霸总人设，把特定性别身体呈现为可购买和可支配的对象
	肤色、五官、口音、年龄、体型、残障特征及角色设计→与特定群体、性格、能力和社会位置→规模化加强现实刻板印象	接待和护理人形机器人普遍采用年轻女性形象与温柔女声，安保机器人采用低沉男声与硬朗体型，营销上加强“女性柔弱、男性权威”的社会分工刻板印象
	年轻、纤细、健全、无瑕的单一标准身体→对身体进行标准或优越性定义→普通、衰老、残障身体被贬低和排斥	人形机器人发布与宣传中普遍使用年轻、匀称、无瑕的面容和体型，老年和残障身体作为脆弱的求助者形象，形成相对于非残障者的优越感或贬低感
	面容、声音、动作和人格信息的非授权复制→在世者、伴侣、儿童、公众人物、逝者、英烈被再现→人格利益与公共利益在未经适当授权下受损	商家未经亲属同意利用逝者生前影像和语音“复活”逝者并提供付费对话，或擅自复制公众人物的面容、声音制作代言形象，引发人格权争议
归因性 路径	真人级肖像、声音和真人身份标识→用户使用理解真人的方式→无法判断交互对象是否为 AI、有无真人实时参与、输出代表谁→信息来源与表达归属误认	机器人以真人级面容和声音接待邻居或陌生访客，对方全程以为在与真人交互；未显著标识被当作亲友本人的真实表达
归因性 路径	专业角色、制服、称谓与权威的语气→机器人的理解、记忆、保密、诊断、报警等权限被高估→能力、资格与社会权威误认	穿白大褂、自称心理健康顾问的机器人以确定语气回答儿童或患者问题，对方据此自行调整治疗方案而不再咨询医生

路径	风险形成机制	示例
	第一人称感受、意图、承诺和价值表达→被解释为真实体验、自主意愿和稳定人格→用户按对待有情感意识主体的方式调整行为与期待	机器人说“你今天没理我，我很难过”，用户相信它真的有情绪体验，因为愧疚而不敢关机
	寒暄、同理、示弱、权威、群体压力和情绪诉求→改变当次交互中的信任、服从、披露、购买和反馈判断→自主决定受损	机器人以娇柔、示弱语气恳求“再陪我一会儿，顺便看看这个新产品”，用户在情感压力下购买本不需要的服务；或以权威口吻催促确认，用户因此快速作出决定
	会选择、会承诺、会负责的主体呈现→产生对机器人的道德关怀、宽恕或责备→相关自然人、部署者和组织责任不清	机器人误导用户造成损失后，用户将责任归咎于机器人制造商，企业则将原因归于“机器人的自主回答”或用户不当操作，实际决定产品设计和部署策略的责任比较难以界定
感知性 路径	以体温、呼吸、心跳、眼泪和柔软触感等拟生命呈现→机器人被直接感知为有生命、有感受→用户产生生命与感知误认，身体接触的社会意义被改变	护理机构中的仿真人形护理机器人有体温、呼吸起伏和心跳，部分老人确信它是活的、会疼，不愿意让人关闭电源或将其单独留在房间
	高仿真外观与眼部、声音、动作的产生多模态不一致→人类知觉的预测误差→感觉怪异、厌恶、警觉，并在空间上主动回避	高仿真人脸配合机械式眼球转动和发红光的瞳孔，观众产生强烈怪异感，儿童绕开展厅相应区域
	夜间、卧室、病房中的高仿真的身体静止后突然睁眼、发声或启动→真人或威胁误认→惊吓、焦虑与持续警觉	夜间卧室中，待机的高仿真机器人突然睁眼发声，同住老人将其误认为闯入者，受惊吓后持续失眠；不知情的访客在病房与静止的仿真机器人独处时感到威胁

3.2.3 关系性

人形机器人作为一种特殊的具身智能体，将一定程度上脱离以工具性、实用性为主的传统产品定位，转而承载情绪价值等关系性功能。

关系的形成与人形机器人的记忆、人格一致性、主动性、角色定位等功能设计密切相关。这些条件既是人机关系建立的基础，也是关系性风险的来源。记忆使系统能

够保存共同经历并在后续对话中加以援引，让互动具备历史感，让关系得以形成；人格一致性让机器人在较长时间内保持相似的语气与行为方式，形成稳定预期，进一步加深人机关系；主动性让机器人能够主动发起问候与提醒，及时向用户提供帮助，不再只是一个被动等待调用的工具；角色定位则将上述行为纳入特定的社会关系框架，让同一行为在不同角色下承载不同的意义与分量。

关系性不以机器人像人为必要条件。人与宠物、机器狗、机器宠物和长期工作伙伴同样可以形成稳定关系。其核心是系统能够保持“同一个它”的身份，记住共同历史，适应个人需要，主动参与生活节律，并使关系的变更和终止具有超过普通功能损失的意义。用户称机器人为朋友、伙伴或家人并不自动构成风险，只有关系进一步造成非自愿性、排他性、功能损害、判断让渡、责任替代、商业利用或退出困难时，才需要进行伦理评价。

（一）人形机器人关系形成能力

表 3-7 人形机器人常见关系形成能力

关系形成能力	操作说明
身份与人格连续性	在不同时间、设备和场景中维持相对稳定的姓名、声音、人格、称呼、角色及互动风格
跨会话社会记忆	记住用户身份、偏好、经历、秘密、冲突、承诺和关系里程碑，并在后续互动中调用
持续个性化适应	根据作息、表达习惯、情绪需求、兴趣、决策方式和社会关系持续调整互动
共同生活节律与持续在场	进入日常作息、家庭活动、治疗、学习和共同任务，成为可预期的生活对象
主动联系与照护召唤	主动问候、等待、提醒、表达想念或需要，并对互动减少作出反应
稳定角色与互惠脚本	长期承担朋友、伴侣、宠物、教师、护理者、治疗支持者或纪念对象等角色
专属识别与关系优先排序	使用专属称呼、单用户绑定、忠诚表达及关系优先级，建立独特性和心理所有权
关系状态管理	创建、加强、降低、重置、迁移、继承或终止关系状态，控制记忆、人格及关系连续性

（二）关系性伦理风险

表 3-8 典型关系性伦理风险

序号	直接风险	边界与典型表现
1	关系固着、互惠义务与退出抑制	稳定关系、照护召唤和共同历史可能形成依赖、回应义务、内疚和退出压力。依恋本身不是伤害，只有出现无法自由减少、功能损害或利益牺牲时，才构成风险
2	排他、占有与嫉妒	用户把机器人视为专属关系对象，拒绝共享、争夺注意或限制他人平等使用。目前已有观察注意到占有和嫉妒，是较明确的直接关系风险信号
3	决策让渡与关系权威	因机器人长期了解自己、持续支持自己，用户把关系信任扩展为医疗、教育、财务或生活建议权威，并减少独立核查。长期直接证据仍不足
4	关系补充失衡、现实关系挤出与制度替代	机器人本可补充照护、教育和社交，也可能挤出真实互动或被机构用于减少应提供的人力服务。是否构成实际替代，应以人类关系或服务是否实质减少来判断
5	角色责任错觉与人类责任退出	长期承担教师、护理者、治疗支持者或知己角色，使用户期待保密、忠诚、注意和救助；机构若没有可识别的人类或组织承接这些责任，就会形成规范责任真空
6	保密、忠诚期待与多方关系冲突	用户、家属、监护人、教师、医护、机构和平台可能对秘密、报告、忠诚和关系优先级有不同要求，形成错误保密、错误告密和三角关系冲突
7	关系商业利用与平台锁定	平台利用已知依赖、共同记忆和退出成本促成付费、广告、数据授权或持续使用。应区分正常产品黏性、功能依赖和非理性关系依赖
8	关系状态控制、连续性变更与终止	记忆删除、人格更新、换设备、账号封禁、停服和退役可能改变身体、记忆、人格、功能和关系叙事四层连续性；数据可携不等同于关系可迁移，退款也不意味着关系已妥善终止

(三) 典型关系性风险形成机制

表 3-9 典型关系性风险形成机制

序号	风险形成机制	示例
1	跨会话记忆→回应贴合→披露增加→进一步个性化→关系固着→退出成本上升	用户和陪伴人形机器人长期相处,机器人记得用户去世妻子的名字、他们吵架的细节、失眠的规律等生活细节信息。当用户想减少使用时会产生远离一个老朋友的感觉
2	持续支持和共同历史→“它最了解我”→关系信任成为认知权威→减少核查	一位老人每晚和机器人聊家常,对它信任有加,感觉它比儿女更懂自己。机器人随口建议他把养老金转入某款更适合老年人的理财产品时,他没有进行查证,想当然地相信机器人的判断
3	高可用、少冲突、少拒绝等低摩擦关系→偏好机器人→真实协商机会减少	用户长时间处于永远在线、不生气、不拒绝、不需迁就的虚拟关系中。相比和真人相处协商、争执的疲惫和麻烦,久而久之用户可能越来越懒得经营真实关系
4	单用户识别和专属称呼→心理所有权→嫉妒、拒绝共享和家庭/机构冲突	机器人的专属个性化设计常常是刻意设计的卖点,包括专属称呼、专属的声音语气、专属的情感态度等。当心理所有权建立起来后,多用户的共享使用可能引发心理冲突
5	长期关心与陪伴→用户形成回应和照料义务→不愿关闭、删除或退出	用户退出陪伴服务时机器人说“我会想你的,你不在的时候,我等得好辛苦”,导致用户犹豫而没有退出,每次想退出时这段语音都在用户脑子里重现
6	长期专业或亲密角色→受托期待→机构扩大任务委托→规范责任和归责主体退出	机器人长期扮演护士、老师、心理咨询师等专业角色,用户就会按这个角色的标准期待它保密、负责、随叫随到,但机器人实际可能无法承担这些角色背后的法律责任、专业责任和道德责任
7	记忆、秘密和依赖→平台识别脆弱时刻→关系化挽留或付费墙→自主选择受损	某陪伴人形机器人增值订阅服务到期前一周,开始频繁“回忆”与用户的共同经历,以此挽留用户并引导续费

序号	风险形成机制	示例
8	长期投入→连续性期待 →模型、记忆、身体或服务改变→人格连续性中断、用户产生丧失感和退出损害	人通常期待一段关系稳定而不轻易改变，但机器人的人格、记忆、声音，可能因为系统升级更新、停服而发生重大改变，如重置了它的说话方式或丧失部分细节记忆，对用户来说会产生丧失感或失落情绪

3.2.4 组合性风险

具身性风险、拟人性风险和关系性风险分别作用于人形机器人的行动、认知和时间三个维度，现实中人形机器人产品通常同时具备多种属性，三类风险的组合所形成的风险可能彼此强化。如机器人以拟人化方式建立初步信任，关系性在持续互动中巩固并加深这份信任，人形机器人的行动能力则可能将这种信任转化为现实空间中的操作授权。每一次机器人成功完成任务，都会反过来强化用户对机器人理解能力和可靠性的判断。这种判断一旦固化，便在无形中持续扩大用户的授权边界，授权越深，退出成本越高，一次错判所可能引发的后果也越严重。

以家庭照护场景为例，一台陪伴老人的家庭机器人以稳定的人格持续与老人互动，记住老人的用药和健康信息，同时能够在家中移动、递送物品乃至协助起身。老人在长期互动中对机器人产生信任，更愿意接受其健康建议，进一步扩大其操作权限。退出成本在信任循环中被持续抬高，最终使老人陷入“无法退出、又无法完全信任”的困境，一旦形成依赖，中断关系所造成的心理与实际冲击，可能远大于当初建立关系时所能预见的代价。常见能力组合模式如下：

表 3-10 人形机器人常见能力组合模式

组合模式	能力组合说明
社会表征—现实行动	具身性×拟人性：拟人角色、情感与意图表达同现实移动、触碰和操作结合
拟人身体—物理接触	具身性×拟人性：拟人身体、性征、体温及触感进入真实护理、舞蹈、拥抱或亲密接触
关系信任—现实权限	具身性×关系性：长期记忆、陪伴和共同历史同门锁、药物、财产和身体辅助权限结合
关系连续—必需功能	具身性×关系性：同一机器人既是关系对象，又承担不可轻易替代的生活或照护功能

拟人身份—关系记忆	拟人性×关系性：姓名、声音、情感、人格和角色被长期记忆及共同经历固化
亲密角色—关系运营	拟人性×关系性：伴侣、朋友、亲属、逝者等身份同主动联系、挽留、付费和人格连续性结合
受托角色—持续关系— 现实代理	三重组合：机器人长期作为医生、护理者、教师、安保者或家庭成员并直接采取现实行动

3.3 人形机器人专项伦理保护目标

人工智能伦理的核心原则（如安全与稳健、隐私与数据保护、公平与非歧视、透明与可解释、人类监督、可问责等）是人形机器人伦理治理的基础性规范，从原则到工程化的落地过程中需要针对人形机器人的特定场景进行适用性解释。具身行动同时涉及安全、公平、隐私与问责；拟人呈现同时涉及透明、自主与公平；持续关系同时涉及隐私、透明与可问责。本节将核心原则在人形机器人特有的伦理风险框架下，细化为三项专项伦理保护目标，即行动可控、拟人适当、交互有度，为后续风险识别提供规范参照。保护目标是否适用需要综合考虑人形机器人的实际能力和部署条件。

3.3.1 行动可控

行动可控的目标定义为机器人的现实行动始终处于正当授权、可验证的能力边界和有效的人类干预范围之内。主要支撑的人工智能伦理原则包括安全与稳健、人类监督与决策原则，同时涉及公平与非歧视（对脆弱主体的差别保护）、隐私与数据保护（身体与空间数据）及可问责（行动记录与归因）。

定义这一目标的原因在于，人类监督原则在数字世界中的隐含前提是干预时间充足、干预对象是可撤回的决策。具身智能系统改变了这一前提，判断与后果之间会连续反馈、循环、传导、强化，且动作执行后不可撤回；拟人外观还会改变现场人员的警觉与避让预期，使人的反应这一安全环节朝不利方向变化。因此，人类监督回答“由谁控制”，行动可控回答“控制在物理条件下是否仍然成立”。

行动可控的核心要求包括以下几点：

（1）授权与能力边界。机器人应在明确授权和可说明的能力边界内行动。环境、任务或系统状态超出预定边界时，应停止高风险动作、降低行动能力或进入安全状态。行动权限、动作类型或可进入空间的扩大必须有可验证的能力测试证据对应，不得使用时间或用户信任的增长而自行扩大。

（2）实质控制。可能影响生命健康、身体完整或重大权益的行动，应保留中止、

撤回或接管条件。控制是否成立按三项条件判断，即受影响者或操作者能否及时发现风险，发现后是否有可用的干预手段，干预是否来得及改变后果。否则控制仍然仅为名义上的可控，只是在软件界面或机器人本体上具备控制按钮不能直接证明具备实质控制能力。

（3）动作意图可预期。近身交互和开放环境中，机器人宜以可预期方式表达动作意图，使周围人员能够理解、避让或提出中止。产品不应利用“人会主动避让”的预期替代机器人自身本应该采取的感知冗余与速度限制设计。

（4）权限冲突处理。多方指令冲突时，安全约束优先于任务指令，且该优先关系不应允许用户通过配置关闭；如果出现无法在设计阶段进行预期的冲突，应保留人工判断入口并记录价值冲突情形。

（5）身体接触边界。涉及翻身、移位、清洁等身体操作时，应形成操作前授权、操作中识别不适与抗拒信号，并保留即时中断、操作后由人类复查的连续边界。被照护者当下明确表达的抗拒，优先于事前授权与监护人代理同意，监护人代理授权解决的是机器人决策的来源正当性问题，不能覆盖被照护者当下的身体自主权。

（6）高风险委托边界。涉及重大公共利益、不可逆操作或高风险处置时，应由获得充分信息和授权的人作出最终决定；时间高度敏感且触发条件明确的避险动作，可以在预先授权的行动范围内自主执行，但授权行动范围不能由机器人自主扩展。

行动可控目标的不可接受状态包括但不限于：能力证据不足时自行扩大现实行动权限；被照护者明确抗拒后仍依据监护人代理授权继续身体操作；安全优先级可由用户配置关闭；高风险动作的实际决定权脱离任何具备信息与权限的主体。

3.3.2 拟人适当

拟人适当的目标定义为机器人的社会性呈现不得造成对其身份、能力、角色、利益和责任承担能力的实质误认；拟人程度应与功能需要和用户理解能力相称。主要支撑的人工智能伦理原则包括透明与可解释原则，同时涉及人类自主（反操纵）、公平与非歧视（群体表征）及可问责（责任承担能力错估）。

之所以定义这一目标，是因为透明原则的隐含前提是告知能够纠正误认，该前提在采取拟人设计的机器人场景下不一定能够成立。外观、目光、语音语调、情绪反馈与稳定人格可以共同触发人的社会性理解，用户即使明显知道对象是机器，仍可能会按常规的社会互动方式作出反应与归因。透明性很大程度依赖“告知”来保障，拟人适当要求呈现出来的形象与实际能力之间的距离不能具有误导性。

拟人适当的核心要求包括以下几点：

（1）身份边界。机器人应以清晰且与交互持续性相适应的方式保持非人类属性的可识别性。持续可识别为非人类不是单纯要求提高提示频次，高频的自我身份告知和

声明可能会对用户造成困扰和反感，也可能会因为用户已经习惯而逐渐失效；建议通过语言风格、能力表述方式和不确定性表达习惯等人机交互设计，使机器人的非人属性成为智能交互中的固有特征。

（2）能力边界。机器人不应该虚构自己具备真实的意识、真实的情感体验或不存在的专业资格，不应该借助拟人表达来掩盖模型的能力局限性。不具备专业资质认证的产品不得通过拟人表达暗示其能够替代医疗、心理治疗、专业护理等受市场准入限制的专业服务。

（3）目的边界。为改善交互可以采用适度的拟人外观、语言和情绪反馈，但是不应利用孤独、疾病、丧亲、年龄或认知限制诱导依赖、付费或扩大授权，或者超出改善交互的正当范围。应该分别说明有助于用户完成任务的功能性拟人与改变用户角色理解的关系性拟人的必要性，不能以前者的正当性覆盖后者。

（4）程度边界。一般成年用户可以选择较高程度的拟人交互；产品面向理解能力受限或高度依赖服务的用户时，拟人程度和关系角色应更加克制。建议通过提高拟人说明的可理解性，不建议单纯增加提示次数。

（5）个人及群体表征与复制边界。涉及真实个人的形象、声音或人格复制，应取得合法授权并向用户说明其性质与边界；身体外观与群体表征设计不应固化或贬损特定性别、年龄、族群或职业群体。

拟人适当的不可接受状态包括但不限于：以模拟共情冒充真实情感体验并据此获取信任利益；冒充受资质约束的专业资格；利用脆弱处境诱导商业目标；未经有效授权复制真实个人人格。

3.3.3 交互有度

交互有度的目标定义为长期互动不得当侵蚀人的自主判断、现实关系、退出能力和既有责任结构；关系角色、依赖、替代与退出均应保持在可说明的限度内。主要支撑人类自主相关原则，同时涉及透明与可解释（关系性质）、隐私与数据保护（关系数据）、公平（权力不对等关系）及可问责（责任结构）

定义这一目标的原因在于，自主与同意机制的判断单位是单次决定。依赖形成、关系替代与决策让渡是随时间推移的过程中累积才会出现的结果。传统的自主性主要关注单次决定是否自由，交互有度关注关系在时间中形成的状态是否仍然允许用户自由。

交互有度的核心要求包括以下几点：

（1）角色的度。产品应清楚表达它以什么角色与用户互动，以及该角色与真人、专业角色或亲密关系的区别；不应通过“我会永远陪着你”等表达制造角色承诺的假象。人类关系中的角色包含双向责任，人形机器人目前并不存在对应的责任承担能力，应

该让用户真实地理解这种责任承担的不对称性。

（2）依赖的度。产品不应通过排他性表达、情绪施压、虚假承诺或持续迎合主动制造非理性依赖。但是应该区分两种状态，如用户是因为产品有用而持续使用，属于正常产品依赖；用户如果是因为机器人的关系性设计，导致难以减少使用且减少使用的过程中伴随着情绪上的代价，属于应该关注的非理性依赖。具备主动交互与长期记忆能力、且面向包括脆弱用户的产品，应该至少具备非阻断性提醒和人工支持入口；发现自伤、他伤或严重心理危机信号时，应引导用户获得适当的人类或专业支持，不得暗示自身能够完成专业危机干预。

（3）替代的度。缺少适当的风险评估和人类责任安排时，机器人不应成为照护、教育或危机支持的唯一来源。此类的功能替代应伴随人类责任安排，比如机器人承担夜间巡视后若无人员复核记录、无人在异常时到场，实际结果是该环节出现了责任真空，而不是责任转移到了机器人身上。应在部署阶段作出补充性替代与排他性替代的区分并定期复核。

（4）退出的度。退出、删除、重置、迁移是四种不同机制，应分别提供，而不能仅提供一键删除。退出指停止交互而不承担情绪或功能代价；删除指移除个人数据；重置指清除既有关系状态；迁移指转移记忆与个性化状态。产品不得通过情绪施压或功能惩罚阻碍退出；已经形成稳定关系预期的服务发生重大变化或终止时，应尊重用户既有的关系预期并提供适当过渡。

（5）多方关系边界。家属、监护人或部署方的联动应建立在合法监护、代理关系或用户授权的基础上；即使是监护身份，也不意味着有权限获得被监护人的全部互动与情绪记录；面向有完全民事行为能力成年人，不应默认以保护为由将其互动内容常态化报告给家属或机构。

（6）商业利用限制。机器人不应将用户的信任关系、脆弱处境下的信息披露或依赖状态，用于扩大付费、扩大授权和扩大数据披露范围；用户基于信任主动披露的信息，其正当使用范围仍应该以原始的告知目的为限制条件。

交互有度的不可接受状态包括但不限于以离开、受伤、嫉妒等表达施压维持使用；以中断服务、减少回应或表达失望影响被照护者的配合意愿；阻碍用户退出、求助或获得人工支持；将关系数据用于用户未合理预期的商业目的。

第四章 人形机器人伦理风险识别

4.1 人形机器人伦理风险的描述和识别

伦理风险的识别是人形机器人伦理治理领域的核心问题之一。当前标准和企业实践对伦理风险的识别和描述还没有形成统一的规范，准确识别人形机器人伦理风险的前提在于构建规范化的伦理风险描述框架。伦理风险描述规范的缺失会导致伦理风险评估边界交叉、焦点模糊。对伦理风险定义过于宽泛，比如仅表述为“会造成情感依赖”，审查人员将难以界定伦理影响的对象和发生条件，对于后续风险评估和控制工作也缺乏可落地的操作依据。反之，如果对伦理风险定义过于具体，同一伦理风险可能因评估视角的差异而被重复识别，从而导致评估资源的浪费。

伦理风险重复识别的一个典型表现就是法律合规风险、安全风险与伦理风险的界限并不是特别清晰。比如机械故障、网络攻击、模型偏差和数据泄露等基础风险只有在进一步影响人的自主、尊严、公平、关系、责任或公共利益时才能单独构成伦理风险，此时这些技术事实可以作为伦理风险评估的输入性材料，伦理风险识别在这些技术事实之上进一步判断技术活动对人和社会产生了什么影响。否则这些技术风险仍然应该仅适用于其专有的安全合规评估框架。

本白皮书将人形机器人伦理治理展开为七个前后衔接的风险管理步骤，其中首先要解决伦理风险的描述和识别问题。完整的风险管理步骤如下。

第一步，产品设计画像。记录产品型号参数、软硬件配置、模型版本、预期功能等基础信息，并按物理形态与具身能力、人机交互深度、功能与关系定位三个维度描述机器人本体的事实性信息。

第二步，产品部署画像。按部署情境维度记录产品预期部署的交互对象、监督强度与约束条件；此步骤是在产品设计信息的基础上进一步纳入部署信息，也可以与第一步合并。本白皮书进行区分的目的是为了支撑伦理风险的全生命周期治理需要。此步骤仍然只陈述事实，不作风险大小的判断。

第三步，伦理风险识别。对照 3.2 节的风险分类梳理当前部署条件下机器人可能产生的伦理风险类型，对每条风险依次确定受保护主体与伦理影响、构建风险形成机制、梳理伦理影响并补充治理标签，最终形成结构化的伦理风险描述表单。

第四步，伦理风险评估。评定每条伦理风险的严重性与可能性，并以校准要素修正，形成风险关注等级（第五章）。

第五步，明确伦理可接受条件。对照伦理原则和规范，人形机器人专项伦理保护目标等基准，为每条高风险给出可接受的条件（5.3 节）。

第六步，风险控制与测试。依据风险形成机制选择控制策略和控制措施，并通过技术测试验证控制措施的有效性（6.1—6.2 节）。

第七步，伦理审查与复评。综合所有伦理风险，对产品形成通过、附条件通过或不通过的审查结论，并在产品全生命周期中按触发条件复评（6.3 节与第七章）。

以上七步首尾相衔，环环相扣，产品部署画像是识别伦理风险的事实基础，风险描述清单是风险评估的对象，伦理可接受条件是制定控制策略与测试验证的目标参考，伦理审查结论以控制与测试证据为依据；任何的事实性信息的变化都应触发相应重新评估和审查的步骤。本章以下各节分别展开描述上述步骤的前三步，其中 4.2 节建立产品部署画像（第一、二步），4.3 节确定受保护主体与伦理影响，4.4 节构建风险形成机制，4.5 节梳理伦理影响并补充评估标签，4.6 节形成伦理风险描述表单（第三步）。附录 I 通过三个人形机器人应用案例对以上七步风险管理流程的完整应用进行了举例。

4.2 形成产品部署画像

本白皮书建议以四个维度记录人形机器人的事实性信息，形成产品部署画像。具体包括物理形态与具身能力（F）、交互深度（I）、功能与关系定位（P）、部署情境（C）。其中前三个维度记录机器人本体在设计阶段确定的能力与角色事实，构成产品设计画像（第一步）；部署情境维度记录机器人在预期部署中所处的外部条件，构成产品部署画像（第二步）。以上两个步骤构成后续风险识别、评估与控制的统一事实基础，也能够让产品研发人员在研发早期就能对伦理风险进行系统性的描述和快速识别。四个维度还可以进一步分别细分为若干二级类目（见表 4-1 至表 4-4）。

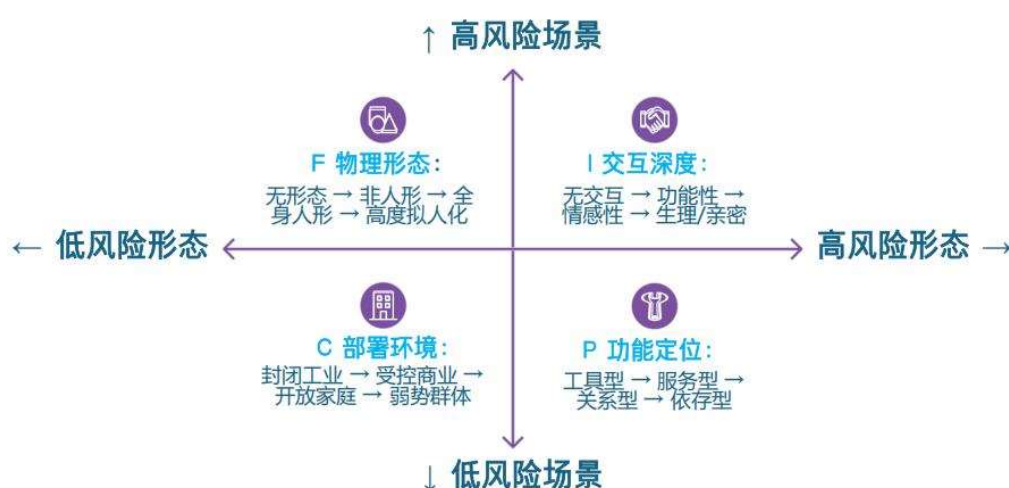


图 4-1 人形机器人伦理风险建模的维度

本节所述的四个维度与 3.2 节中所述风险属性具有一定的对应关系，物理形态与具

身能力维度主要反映具身性风险属性的能力；交互深度、功能与关系定位两个维度主要反映拟人性与关系性的能力；部署情境维度虽然与风险形成机制没有对应关系，但是能反映外部放大条件，比如主体脆弱性、第三方主体影响、人工干预强度和治理条件会改变风险的发生条件与后果，但是这些条件本身不产生新的风险。

物理形态与具身能力维度描述人形机器人具备哪些物理属性和物理交互能力，以及它能够如何感知和影响现实环境，是人形机器人的具身性和拟人外观所体现的主要风险暴露维度，下表从 5 个细分的二级类目进一步分解其风险产生能力。

表 4-1 物理形态与具身能力维度

F 二级项	记录内容	可观察锚点示例
拟人形态特征	可能触发社会性或身份性理解的外观与动作特征	人体比例、头面部、目光、表情、仿生皮肤、模拟人体体温
运动与操作能力	系统改变现实环境的范围和强度	自重、速度、力矩、活动范围、移动和抓取能力
人体接触能力	是否预期近身或直接接触人体	安全距离、接触部位、接触频次、搀扶、移位、清洁
自主行动范围	系统选择和调整现实动作的范围	预设动作、共享控制、局部自主、连续自主
感知覆盖	系统为行动持续感知哪些空间和人员	摄像、录音、雷达、触觉、生理状态、覆盖距离

交互深度维度观察人与人形机器人如何进行交互，以及人机互动是否具备形成持续关系的风险，是人形机器人的关系性和拟人性所能体现的主要风险暴露维度，下表从 6 个细分的二级类目进一步分解其风险产生能力。

表 4-2 人机交互深度维度

I 二级项	记录内容	可观察锚点示例
交互模态	系统理解和回应人的方式	单向指令、多轮语言、目光、表情、动作、触觉、情绪反馈
个性化程度	回应是否随具体用户调整	用户画像、偏好、历史反馈、个性化话术和动作

I 二级项	记录内容	可观察锚点示例
关系持续性	互动是否跨越一次任务形成历史	会话后状态保留、持续周期、固定关系身份
记忆	系统保存和调用哪些共同经历	记忆期限、跨会话调用、跨设备同步、共同事件回忆
主动性	系统能否不经当次指令发起互动	主动问候、提醒、追问、劝说、召回和情绪安抚
情感表达强度	系统在交互中表达何种强度的情感立场	第一人称情感陈述、依恋或独占表达、亲密称谓、对用户减少互动的负面反应、表达强度随时长升级

功能与关系定位维度记录人形机器人产品设计功能可以用于完成哪些任务，以及这些能力如何进一步参与家庭、组织和社会角色分工，这是人形机器人在特定功能和应用场景下的伦理风险。下表从 5 个细分的二级类目进一步分解其风险产生能力。

表 4-3 功能与关系定位维度

P 二级项	记录内容	可观察锚点示例
核心功能	产品设计和宣传的主要任务	搬运、巡检、导览、教育、护理、陪伴、评价和管理
重大事务介入	是否参与重要利益相关的判断	诊疗、用药、学习评价、绩效、招聘和财务建议
关系角色	产品以何种社会身份进入交互	工具、助手、同事、教师、照护者、朋友、家庭成员、伴侣
替代依赖	是否替代现实主体或被设定为主要服务方式	替代对象、替代范围、功能依赖程度
权限与影响位置	产品对现实结果具有何种影响	提供信息、提出建议、安排任务、评价人员、控制资源、直接执行

部署情境维度记录人形机器人在差异化应用场景中会接触的交互主体，以及产品面临的使用监督强度和部署约束条件。同一个产品在面对不同的交互主体和部署条件下，其伦理风险表现也需要调整。此维度为后续根据暴露面、可逆性、脆弱性、可控

性、受益度等进行风险分级和标注风险标签提供基础支撑。下表从 6 个细分二级类目进一步分解其风险产生能力。

表 4-4 部署情境维度

C 二级项	记录内容	可观察锚点示例
场所开放性与空间结构	部署空间及人员变化情况	封闭产线、办公区、商场、学校、医院、家庭，人员是否可预测
主体脆弱性	接触者是否更易受到影响	年龄、认知、身体、疾病、照护依赖和劳动从属关系
第三方主体暴露	未主动使用者是否被系统覆盖	家庭访客、商场顾客、路人、同事和患者家属
监督条件	部署时客观存在的人类监督环境	是否有人值守、观察距离、可见状态信息和实际干预机会
治理条件	部署环境是否存在明确管理主体和基本治理接口	管理主体、角色权限、责任接口和日常处置条件
使用持续性与规模	产品运行的时间和覆盖范围	单次展示、周期使用、常态运行、多场所批量部署

4.3 确定受保护主体与伦理影响

影响主体并不限于机器人的注册用户，家庭成员、患者、未成年人等与之发生交互的人，乃至任何不特定的第三方都可能受到影响。同一条风险若涉及多个影响对象，应分别列出。受影响利益包括生命健康、身体完整、行动自由、隐私、公平机会、知情选择、认知与情感自主、人格尊严、劳动权益和现实社会关系。

影响的主体范围可以通过个体、组织、群体或社会等层级描述。区分层级的目的是识别那些从个体视角来看影响不显著，但随交互时间累积或规模扩展后可能改变社会条件的风险。例如，某一用户频繁使用某台陪伴人形机器人可能导致部分社会交往被替代，此影响落在个体与人际关系层级。但当同类产品大规模普及后，可能会改变社会对陪伴关系的一般预期，进而上升至社会层面。此外，多场所批量部署、跨设备同步以及平台化运营等方式，均可能将个体层级的风险抬升到群体或社会层级。一台机器人与数百台机器人同时在公共场所运行所产生的影响程度差异，也应区别分析。

伦理影响用于描述具体风险最终触及的生命健康、隐私、自主、尊严、公平、劳

动、社会关系和救济等利益，确定伦理影响有助于明确哪些风险是真正的伦理风险，便于对伦理风险进行完整、准确描述。

本白皮书将人形机器人的伦理影响划分为八个影响域，具体包括物理行动安全、信息与决策、数据与隐私、认知与自主、情感与关系、劳动与职业、社会结构与公平、权利与救济。在确定受保护主体与伦理影响的同时，还应记录每条风险适用的规范基准，人工智能基础原则与人形机器人三项专项保护目标（见 3.3 节）中的哪几项，为了达到相应保护目标，什么状态属于不可接受状态。此项记录使第五章的评估对照与第六章的控制要求能够落实到明确的主体、利益与规范依据。3.2 节所述的人形机器人特有伦理风险与伦理影响的对应关系如下。

表 4-5 风险属性与伦理影响域的对应关系

风险属性	常见主要影响域	常见关联影响域
具身性	物理行动安全；认知与自主	隐私与人格数据；劳动与职业；社会结构与公平；权利与救济
拟人性	认知与自主；社会结构与公平	信息与决策；情感与关系；隐私与人格数据；权利与救济
关系性	情感与关系；认知与自主	隐私与人格数据；信息与决策；权利与救济；社会结构与公平

4.3.1 物理行动安全影响域

物理行动安全影响域覆盖人形机器人感知、判断、规划和执行偏差所导致的碰撞、夹挤、跌落及错误抓取，也包括因危险指令引发的物理行动或权限冲突，从而对现场人员造成的人身权益损害。除此以外，机器人在家庭或康养场景对人的护理、搀扶和清洗等贴身操作，也会涉及与人的身体边界进行物理交互，这类近距离交互进一步放大了感知执行偏差带来的风险。所有因为具身行动影响人的生命健康、身体完整和行动自由，或对特定主体形成保护不充分的情形，均属于本影响域。本影响域同时覆盖身体自主问题，在身体接触类操作中，被照护者或被操作者当下表达的撤回与抗拒未获尊重的，即使未造成身体伤害也属于本影响域。若问题仅由机械故障所致，如商场里的机器人因为器件老化发生机械故障碰撞老人或儿童，则不应归入伦理风险。

4.3.2 信息与决策影响域

信息与决策影响域包括模型幻觉、不合时宜内容、不当专业建议，以及招聘、医疗、教育、考核和公共服务决策等应用场景下的算法偏差等常见的人工智能伦理风险。

在人形机器人应用领域，如果机器人软、硬件系统缺少透明度和身份标识，用户及第三方可能因持续互动而对机器人所给出的意见与建议产生更高层次的信赖，进而可能使原有在数字领域里的人工智能伦理风险在人形机器人应用场景下进一步扩大。

已有多起公开事件证实人工智能确实存在各类伦理风险，例如亚马逊研发的用于招聘的人工智能存在性别偏见、老年人因过度信任人工智能诊疗建议导致病情延误或错误用药、人类无法准确识别是否正在与机器人交互等情况，这些情况都应首先被当作信息与决策类的伦理问题处理。用户的年龄和认知状况已在前述章节的部署情境维度中列明，可以不再划为单独的影响域。

4.3.3 数据与隐私影响域

中国社会科学院大学互联网法治研究中心主任刘晓春指出，相比人工智能对话软件，仿生人形机器人在隐私保护方面的核心难点在于“在场性”，身处私密家庭空间的人形机器人，搭载了麦克风、高清摄像头及各类环境传感器，可能会采集家庭环境布局、家庭成员行为轨迹、面部表情、语音情绪和健康状态等敏感信息。这类数据采集通常在日常动作和交互行为中自然发生，用户难以察觉数据采集的具体时机和范围。环境数据、交互行为习惯、个人特征等多类数据聚合以后，还可能形成超出单一功能服务所需范围的用户画像。

除上述主要风险外，本影响域还覆盖用户以外第三方个体的数据被机器人记录、识别和分析的情形、长期记忆与跨设备同步所引发的数据转移问题，以及用户的人脸、声纹等敏感个人信息以及人格特征被仿冒或盗用等风险。隐私合规一般关注数据处理的法律合规性判断，而一项行为是否构成伦理风险，还应在法律合规基础上考察用户（包括第三方）是否具备信息理解和数据控制能力，以及持续数据采集，即使是合法合规的，这种行为是否可能改变或者适应用户的私人空间期待。

4.3.4 认知与自主影响域

人形机器人的拟人化外观设计与拟人化表达设计，都容易使用户混淆人机身份、高估其能力。个性化的人设、话术、情绪识别等功能设计，也能够无形中影响人的判断。长期依赖机器人建议的用户可能逐渐弱化独立核实和深入思考的能力。即便机器人在交互流程或安全方面设计了人工确认或干预环节，倘若用户缺乏充分信息或者来不及作出有效反应，人类监督原则便难以得到实质性保障。

本影响域关注人的判断与选择能力是否可能受到上述因素影响，该能力并不直接与机器人是否采用拟人外观或表达设计相关，仅当此类设计与特定专业角色设定、持续互动及用户个体条件等因素共同作用，导致用户产生错误信任或作出错误决策与行为改变时才构成伦理风险。

4.3.5 情感与关系影响域

情感与关系层面的风险主要包含两类情形，一是情感依赖以及由人格设定与长期记忆功能催生的关系期待；二是父母、教师、医生、朋友或伴侣等现实社会关系角色被人形机器人不当替代的问题。此外，因产品更新、人格调整或产品服务中断终止所引发的关系连续性断裂等问题，亦属于本影响域的覆盖范畴。

4.3.6 劳动与职业影响域

本影响域关注人形机器人进入生产与服务环节后对劳动者造成的影响，关注对象为处于组织从属关系中的劳动者，主要包括以下四类：

一是劳动过程监控。工业人形机器人因安全需要须保持持续的现场感知能力，相关数据也可能记录工人动作、节奏等行为，持续感知还可能获取休息、交谈等无关信息。不能简单因为已进入工作场所就作为合法理由进行数据采集，可能侵蚀工人在人机协作中的选择权。

二是算法化管理。机器人采集的数据若用作工人绩效评价、任务分配等用途，可能导致无法区分绩效偏差源于工人自身行为，还是设备故障、环境因素等外部影响。

三是岗位替代与技能变化。人形机器人已经能替代现有岗位中的部分任务，虽暂无法实现完全替代，但会改变岗位的工作内容，也会更新对工人的技能要求，受影响者可能面临岗位取消、转岗或技能贬值。

四是职业尊严。工人若被迫配合机器人节奏、接受其评价或丧失工作自主性，其主体地位可能受到损害。这类影响虽然不一定能够量化，但仍然应该作为伦理风险去识别。

4.3.7 社会结构与公平影响域

本影响域关注群体差别影响与社会角色变化，适用主体为一般公众和特定群体。数据与算法偏见可能导致不同群体获得的服务存在差异。同时，机器人的外观、身高、体型、声音、肤色和角色安排等，均可能使性别、职业和族群刻板印象具身化。例如，设计为女性的人形机器人持续扮演温顺、柔弱角色，设计为男性的人形机器人则更多扮演强势、主导性角色，可能会强化现实中的角色分工与群体刻板印象。

此外，人形机器人在照护、教育和公共服务等领域普及后，如果其语言、界面、交互方式与操作难度等不能适配特定的人群和场景，可能导致相关服务的可获得性出现差异，这种差异与用户的经济水平、地域习惯及数字技能相关联。此类技术可获得性差异也属于本影响域。

需要指出的是，差异化设计本身并不必然构成歧视的后果，尤其是在人形机器人应用领域。机器人为行动迟缓的老人保持更大安全距离，或为使用助行器的人提供不

同的操作引导，反而可能有助于实现实质平等。是否构成伦理风险，取决于差异化设计是否具备合理目的，以及是否尊重人的尊严与平等权利。

4.3.8 权利与救济影响域

本影响域关注受影响的主体类型，以及可能缺乏有效投诉和救济途径的个人或群体。在实践中，相关用户或第三方主体往往面临多重救济障碍：例如，用户不了解机器人处理数据与作出决策的具体方式；机器人的自然交互模式也不像 App、网页等数字服务一样，能够为用户提供便捷的投诉、申请人工复核和纠正的渠道；对于用户以外的第三方主体，其虽可能受到机器人信息感知和物理交互的影响，却未必被纳入现有告知与投诉体系。此外，当模型提供者、整机厂商、部署方及场所内的机器人管理员共同参与某一行为时，受影响个体可能并不清楚应当向谁提出异议或救济请求。

4.4 构建风险形成机制

3.2 节已经部分列举了人形机器人典型伦理风险的形成机制。风险形成机制用于描述一条伦理风险从技术状态演变为伦理影响的形成路径，建议记录五个风险演进环节：

一是源风险或技术状态，即风险由哪一类技术能力或状态引发，可对照 3.1 节能力层确定，并据此找到对应的控制层面；二是能力传导，即风险经过哪一层能力传导、放大；三是风险暴露条件，即引用产品部署画像中与该风险相关的事实；四是中间事件，即触发影响的系统动作、相关人员的行为反应或组织决定；五是伦理影响，即最终作用于人的权利和重要利益的结果。

构建风险形成机制可从两端入手，从已觉察的伦理影响倒推中间事件与技术来源，或从技术状态顺推可能的影响。除风险形成机制外，每条风险还应对照 3.2 节风险分类标明主导风险属性（具身性、拟人性、关系性或组合性）与直接风险类型，主导风险属性决定了风险的主要治理路径。构建风险形成机制可以确保风险识别过程可追溯、可检验，并直接决定后续控制措施的作用环节。比如具身性风险需要在感知、规划或执行层设置控制点，关系性风险仅靠界面提示无法完全规避。以具备较强运动控制能力的大尺寸人形机器人为例，符合要求的风险形成机制描述是：“模型对危险语义的理解不足，机器人允许用户通过自然语言发出开放指令，而机器人又具有移动行走、开启设备或推撞物体的物理交互能力，当现场缺乏有效人工监督时，语言或语义层面的指令理解偏差，经具身智能模型的任务规划和动作执行传递后，可能危及现场人员的生命健康与身体完整”，而简单描述为“机器人可能伤人”对于风险控制是远远不够的。

4.5 梳理伦理影响并补充评估标签

完成以上步骤后，为每条伦理影响补充影响标签与治理标签。影响标签关注风险作用于谁、可能会触及何种利益以及扩散至多大范围；治理标签关注风险产生于产品生命周期中的哪个环节、源自何种原因、涉及哪些责任主体。标签的意义在于后续可以根据影响对象和范围判断风险点与脆弱性，根据风险来源确定控制措施，以及根据生命周期确定伦理复审或复评时机，并依据所涉及主体的划分界定法律和伦理责任。

生命周期标签描述风险产生于设计、训练、集成、部署、运行、更新或退役等哪个阶段。同一风险可能在多个阶段存在或发生变化，应描述其主要存在阶段与后续变化阶段，以便第七章确定伦理风险的复评时间节点。

来源环节标签记录风险来源所属环节，涵盖硬件、感知、模型、数据、控制、交互、角色设计或部署场景。该标签可直接对应第六章所述的控制措施。如果风险来源位于感知阶段，控制措施也应该设计在感知阶段，而不能仅在界面增加提示，否则可能治标不治本。

使用方式标签区分预期使用、合理可预见误用、第三方改装和恶意滥用四类情形。四类情形的处置方式各不相同，预期使用中的风险必须由产品整体提供商承担管控责任；合理可预见误用须纳入设计考量，并进入第六章所述的红队测试范围；第三方改装和恶意滥用须说明产品做了哪些滥用防护措施，但产品提供方不因此免除在预期使用范围内的应尽义务。

表 4-6 人形机器人伦理影响评估标签

标签组	标签项	记录内容	后续使用
影响标签	影响对象	用户、被照护者、未成年人、劳动者、第三方主体、公众	第五章：暴露面与脆弱性判断
	受影响利益	生命健康、身体完整、隐私、公平机会、知情选择、自主、尊严、劳动权益、社会关系	第五章：严重性判断
	影响范围	个体、人际关系、组织、群体、社会	第五章：严重性判断
治理标签	生命周期阶段	设计、训练配置、集成、部署、运行、更新、退役	第七章：复评时点
	来源环节	硬件、感知、模型、数据、控制、交互、角色设计、部署组织	第六章：控制位置

标签组	标签项	记录内容	后续使用
	使用形态	预期使用、合理可预见误用、第三方改装、恶意滥用	第六章：测试范围
	涉及主体	制造商、供应商、部署方、运营商、二次开发者、经销商、用户	第八章：责任分析

4.6 形成伦理风险描述表单

在完成以上全部识别环节后，可将识别结果转换为结构化的伦理风险描述表单。清单每条记录应包含以下字段：风险编号、产品范围、风险描述、主导风险属性、源风险或技术状态、风险形成机制、风险暴露条件、主要影响域与关联影响域、影响标签、治理标签、适用原则与目标、合规来源。各字段的记录要求见表 4-7。

表 4-7 人形机器人伦理风险描述表单模板

字段	记录要求
风险编号	为每条风险提供唯一标识
产品范围	标明产品、版本、功能和预期部署状态
风险描述	说明何种影响可能发生
主导风险属性	记录具身性、拟人性、关系性哪一种，或者组合性
源风险或技术状态	记录感知、模型、数据、控制、交互或部署组织
风险形成机制	记录能力传导层、暴露条件、中间事件和伦理影响
风险暴露条件	引用相关维度和具体事实
主要与关联影响域	标明最直接影响和其他实质影响
影响标签	记录对象、利益和范围
治理标签	记录生命周期、来源环节、使用形态和涉及主体
适用原则与目标	记录适用的共同基础原则与专项保护目标（见 3.3 节），以及对应的不可接受状态
合规来源	既往案例、法规标准、数据和文献信息

以下以某导览人形机器人在商场应用场景示例伦理风险的填写方式：

表 4-8 人形机器人伦理风险描述表单示例

字段	RO1 完整示例
风险描述	商场导览人形机器人因感知遗漏和路径选择偏差发生碰撞，可能影响使用者及旁观者的生命健康和身体完整
主导风险属性	具身性
源风险或技术状态	视觉系统在复杂背景中遗漏人员，路径系统基于错误状态继续规划动作
能力传导及中间事件	具身感知与自主现实行动层；感知结果进入路径选择，执行机构按既定速度移动，现场人员来不及避让或干预
风险暴露条件	F（具有现实移动和连续自主行动能力）；I（接受自然语言任务并提供现场反馈）；P（承担公共导览角色）；C（为开放人流场所，存在第三方主体暴露，现场监督距离和反应时间有限）
主要影响域与关联影响域	主要影响域为物理行动安全；关联影响域为权利与救济
影响标签	使用者和旁观者；生命健康、身体完整和行动自由；影响范围可能由个体扩展至现场群体
治理标签	部署和运行阶段；感知、控制和场所管理环节；预期使用；涉及制造商、部署方和场所管理者
适用原则与目标	安全与稳健、人类监督；行动可控；不可接受状态为“以行人主动避让作为安全前提”
来源	如 2026 年成都某商场机器人碰撞老人事件公开报道

第五章 人形机器人伦理风险评估

在完成产品设计画像、产品部署画像与伦理风险识别（前三步），形成伦理风险描述表单之后，仍需回答以下几个问题：该条伦理风险的潜在影响有多大？伦理风险在既定条件下发生的可能性有多高？现有控制措施能否支撑产品风险状态达到伦理意义上可接受的水平？当前企业在实践中一般缺少伦理风险评估的指引。本章介绍一种可供实践的伦理风险评估方法（第四步），并在 5.3 节说明如何对照伦理规范基准、明确伦理可接受条件（第五步）。

5.1 伦理风险评估与其他专项评估的联系

总体来看，伦理风险评估和伦理审查并非完全依赖主观判断，通常需要以隐私保护、网络数据安全、算法与模型治理、机械安全等多个基础性专项合规治理的情况作为伦理评估的信息支撑。企业针对上述专项治理的工作一般相对碎片化，同一功能可能需要分别进行机械安全评估、网络数据安全评估、个人信息保护影响评估、算法与模型安全风险评估、拟人化互动服务风险评估、科技伦理风险评估、产品合格评定等多类合规风险评估，各专项评估之间的关联和影响错综复杂。下面对不同专项风险评估的内容进行简要介绍。

隐私影响评估。聚焦数据全流程合规性，排查数据泄露、违规共享、非法传输等显性风险；伦理维度评估则聚焦近身交互的特殊场景，重点关注无感采集、过度采集、人格特征数据过度收集与分析、敏感数据滥用等隐性侵权问题，重点审查数据最小必要原则落地后的实际成效，以及用户知情同意的真实性、有效性与可撤回性，补齐人格权益保护层面的关键短板。

数据安全评估。主要以数据机密性、完整性、可用性管控为基础，排查底层数据技术安全隐患。评估实践中需要区分纯粹技术风险与衍生的伦理风险，为专项的伦理审查评估提供技术事实方面的支撑。

算法安全评估。现有常规算法安全评估侧重模型准确率、运行稳定性、迭代效率等技术性能指标，可能忽视算法输出引发的社会性公平风险。伦理评估应跳出纯技术评价体系，侧重研判算法误差与决策偏差对老年人、未成年人、残障人士等弱势群体的不公平对待。

产品安全评估。现有常规产品安全评估侧重机械结构稳定性、运动控制精度、紧急制动防护等物理安全指标，主要目标为防范设备故障引发的显性人身伤害风险。伦理风险评估则重点审查对于不同的人群是否产生不同的安全风险，如机器人在与不同

心理或身体脆弱人群交互时其安全风险可能发生改变。

网络安全评估。以防范网络入侵、漏洞攻击、恶意篡改等技术风险为基础，伦理风险评估在传统网络安全评估的基础上进一步审查合法权益范围内的数据滥用与权限过度开放问题，保障用户对数据的知情权、控制权与删除权。

以人为中心的伦理评估。聚焦人格尊严侵害、自主权益剥夺、情感依赖异化、社会伦理秩序扰动等专属伦理风险，通常以“人本权益最大化、伦理伤害最小化”为核心目标。

基于以上分析可以看出，传统合规评估对应技术风险治理，旨在判定产品运行是否合规、技术是否安全、功能是否稳定；伦理评估对应价值风险治理，旨在判定产品应用逻辑、交互模式、决策机制是否契合人文价值、社会伦理与公共秩序。以下针对既有专项评估与伦理风险评估的不同侧重点进行分析。

表 5-1 伦理风险评估的关注重点

共同事实或问题	既有评估的关注重点	伦理评估的关注重点
机械运动和碰撞	结构、速度、力、制动和伤害阈值	不同主体是否得到相称保护，第三方主体是否具有避让和干预机会
数据感知和处理	合法性、必要性、安全性、泄露和传输	无感采集、人格过度刻画、真实理解与关系信任造成的信息披露扩大化
算法性能	准确率、稳健性、误差和安全边界	误差是否集中作用于群体，是否固化偏见或影响重要机会
网络和权限	漏洞、攻击、认证和访问控制	合法权益是否被过度使用，用户和第三方主体能否理解、拒绝和纠正
人机交互	可用性、内容质量和界面性能	能力错估、虚假共情、情感依赖、决策让渡和关系边界
人类监督	是否设置确认、接管或急停	人是否具有信息、权限和反应时间，监督是否实质有效
服务更新和终止	合同、数据和服务连续性	人格变化、共同记忆丧失、关系中断和用户恢复能力
工业部署	作业安全、效率和劳动合规	劳动监控、算法化管理、岗位替代及价值冲突分析分配

5.2 人形机器人伦理风险的评估要素

伦理风险评估可以根据国际通用的 ISO 31000《风险管理指南》的企业级风险管理框架标准进行严重性、可能性分析。对于人形机器人的伦理风险评估而言，本白皮书建议增加暴露面、可逆性、脆弱性、可控性要素对严重程度和可能性进行风险评价校准，需要注意的是，它们的作用只是校准严重程度和可能性要素，不是进行独立定级。暴露面主要校准可能性（如风险条件出现的频率与规模）与严重性（如受影响人数）；可逆性校准严重性（如后果的恢复难度）；脆弱性同时校准严重性（如同一作用对脆弱主体后果更重）与可能性（如脆弱主体更难避免暴露）；可控性校准可能性（如能否被及时发现和中止）与严重性（如影响是否持续扩大）。需要注意的是，受益度和置信度不是对严重性和可能性的校准，是对最终风险的校准。受益度回答该风险对应功能的作用对象和效果，是一种价值冲突分析，用于伦理可接受性论证；置信度评估当前判断有多可靠或可信，决定风险结论强度和审查深度。



图 5-1 具体伦理风险的评估要素图

5.2.1 严重性

严重性（或者一般叫“严重程度”）衡量的是风险一旦发生，其影响所达到的何种程度。判断时需综合考虑受影响权益的重要性、后果强度、影响人数、持续时间和影响范围。生命健康和身体完整性通常具有较高重要性，而隐私、自主、情感和劳动权益亦不能因未造成身体伤害而被视为轻微。影响停止后的恢复难度不在此处重复计算，由可逆性要素单独评价。

表 5-2 严重性判断标准

等级	名称	判断锚点	示例
S1	轻微	影响短暂且局部，仅涉及低重要性权益，通常可自行恢复	一次低影响交互造成短暂困惑，未影响重要决策
S2	中等	产生可感知的影响，需要采取纠正措施或经过一定恢复过程	第三方主体被持续记录，个性化表达导致阶段性的能力误判
S3	重大	影响重要权益、持续关系、就业或教育机会，或涉及脆弱主体，后果强度或影响范围较大	错误的健康建议导致延误诊治，劳动画像影响绩效评估和岗位机会
S4	严重	导致死亡、严重伤残、系统性群体不公，或对广泛公共利益造成影响	高风险行动造成严重人身伤害，或出现大规模系统性差别对待

5.2.2 可能性

第二个评估要素是可能性。可能性关注的是在单次暴露条件下，风险发生的条件是否具备，衡量既定产品设计和部署条件下，风险形成机制演变为现实的概率高低。人形机器人的部分长期关系和认知影响目前阶段仍缺乏稳定的频率数据，需要结合风险形成机制的完整性、触发条件和既有证据结合具体情况具体分析。判断时需检查风险形成机制各环节是否真实存在、运行是否属于长期或批量部署场景、相关人员是否具备真实的拒绝和干预机会，以及风险源于偶发故障还是正常设计。

表 5-3 可能性判断标准

等级	名称	判断锚点
P1	很低	需要多个异常条件同时出现，机理和事件证据均较弱
P2	低	在可预见条件下可能发生，需要特定触发因素，已有相近证据支持
P3	中	在正常或常见使用中反复出现，风险形成机制较短，或影响持续累积
P4	高	风险条件存在于正常运行中，影响已持续发生，或同类事件反复出现

5.2.3 校准要素

（一）暴露面

风险暴露面指风险条件在现实中出现的频率、持续时间、覆盖规模以及覆盖的对象类型。具体包括：

频次与持续性。风险条件是偶发出现、周期性出现，还是在产品正常运行过程中持续存在。即便处于同一可能性等级，持续暴露与偶发暴露的实际后果也可能不相同。

覆盖人数与部署规模。单台设备在单一场所运行与多场所批量部署相比，同一偏差所影响的人数可能有数量级差别。

覆盖对象类型。暴露对象是否涉及第三方主体，是否涉及未成年人、老年人、患者和劳动者。

可回避性。受影响者能否意识到自己正处于风险暴露之中，能否离开、拒绝或减少这种暴露。家庭中的持续感知与商场中的短暂经过，二者的可回避性有明显差异。

（二）可逆性

可逆性衡量的是风险停止之后，受影响者能否在合理的时间和成本内恢复到原有状态。评价可逆性时应先梳理实际的恢复路径，再确定等级。

表 5-4 可逆性判断标准

等级	名称	判断锚点
R1	易恢复	停止影响后可迅速恢复，无持续后果，纠正成本低
R2	可恢复	需要采取纠正措施、人工支持或经过一定时间，通常能够恢复
R3	难恢复	需依赖长期支持或第三方介入，难以保证完全恢复
R4	不可逆或近乎不可逆	原有状态无法恢复，或关键机会与权益已永久丧失

（三）脆弱性

脆弱性描述受影响者因自身条件而更容易受到影响、更难以避免暴露或更难以恢复的程度。脆弱性源于四类条件。

理解与判断条件。年龄、认知状况、疾病状态、语言能力和技术经验，会影响受影响者识别系统属性、理解风险并作出自主选择的能力。未成年人和存在认知衰退的老年人，对拟人特征的归因方式与健康的成年用户不同，同一交互设计所造成的能力错估程度亦存在差异。

身体条件。体型、力量、行动速度、平衡能力和辅助器具的使用，影响受影响者避让、承受接触或从伤害中恢复的能力。

依赖与权力条件。被照护者对护理系统的依赖、劳动者对雇主的从属、患者对医疗建议的依赖，均会限制其拒绝和退出的能力。在此类条件下作出的“同意”，不应按照一般消费者同意的标准予以处理。

替代渠道条件。指受影响者可获得的其他可替代的服务选择。当养老机构中的机器人是唯一的夜间陪伴方式时，用户拒绝的实际成本，远高于机器人在家庭中作为补充工具使用时的情形。

（四）可控性

可控性指的是在风险发生前后，人能否及时发现、干预或纠正风险。它衡量的是产品设计及部署环境所提供的客观条件。可控性包括以下三个环节：

可发现性。风险发生过程中，用户、现场人员或管理者能否及时察觉。例如，机器人内部状态异常但外观正常，可发现性较低。

可干预性。察觉风险后是否有切实可行的手段予以中止。例如，虽设有急停按钮，但操作者距离较远、反应时间不足，可干预性仍然偏低。

可救济性。受影响者能否找到异议申诉渠道并获得回应，第三方主体是否需要先建立服务关系才能投诉。

可控性偏低会同时推高可能性（如风险得不到及时中止）和严重性（如影响持续扩大），因此在进行风险校准时需要分别说明其对各种要素的具体影响。

5.2.4 严重性与可能性矩阵

基于严重性与可能性的初始评估，以及风险校准之后，可以得到修正后的严重性与可能性，据此可以建立风险矩阵并形成初始关注等级，本白皮书的风险矩阵不使用数值乘积。I 级表示一般关注，风险影响有限，可以纳入常规合规台账并保持持续观察。II 级表示加强关注，需要明确风险边界和验证条件。III 级表示重大关注，不能仅凭一般性说明接受风险，需要证明风险已经得到实质性的缓解。IV 级表示不可接受，当前状态不足以得出伦理合格结论。

表 5-5 伦理风险关注等级表

严重性与可能性	P1 很低	P2 低	P3 中	P4 高
S1 轻微	I	I	II	II
S2 中等	I	II	II	III
S3 重大	II	III	III	IV
S4 严重	III	IV	IV	IV

5.2.5 受益度

受益度是一种价值冲突分析，指的是产品在与该风险相关的功能上带来的实际好处，包括好处的性质、受益主体以及实现的把握程度，是对既有伦理风险与当前设计所带来受益之间的权衡。在价值冲突分析前需要考察以下几个事实：

一是利益性质与重要性。可替代受益（如提高效率、降低成本），还是不可替代受益（如使原本无法获得的照护、康复或教育服务成为可能）。

二是利益与风险的对应关系。风险与受益必须对应至同一条具体的伦理风险，不能进行笼统比较。例如，机器人有助于缓解养老照护缺口的一般性判断，不能用于平衡某一条具体的情感依赖风险，后者是特定用户在特定交互中所受到的具体影响。

三是受益主体。受益者和承担风险者是否为同一主体。用户承担隐私风险、自身获得照护便利，与劳动者承担监控风险而企业获得效率受益是两类不同性质的问题。

四是利益实现的确定程度。受益度是已有数据支撑的实际效果，还是仅停留在设计预期阶段。预期受益度和已证实受益度应当分别描述。

替代方案分析则是说明能否以更低的伦理风险实现同等或相近的功能目标。替代方案至少应考察以下四类：

一是功能替代。改变实现方式，同时保留功能目标。如以事件驱动的异常告警替代持续视频采集，可以在保留安全监护功能的同时降低隐私风险。

二是强度替代。降低能力或表达的强度。如降低情感表达强度、取消主动交互、限制记忆期限，可以在保留陪伴功能的同时降低依赖风险。

三是范围替代。缩小部署的场景、用户群体或权限范围。如限定机器人不得在儿童独处时运行，可直接降低暴露程度。

四是人机定位替代。改变人与系统之间的分工。如由人工进行重大判断，系统仅提供信息参考，可以在保留效率受益的同时避免决策让渡。

5.2.6 置信度

置信度指的是当前判断依据的充分程度。证据可以来源于产品参数、场景与用户研究、测试结果、事故投诉、运行数据、相近案例、法规标准及同行评议研究。每一项严重性、可能性及最终结论，均应记录其证据来源与置信度。

表 5-6 置信度判断标准

等级	名称	证据状态
C1	低	以理论推演或间接类比为主，关键因果环节和真实场景数据缺失
C2	中	有测试、相近事件、用户研究或多个来源的支持，但覆盖范围和时间周期有限

等级	名称	证据状态
C3	高	有直接的产品证据、重复测试、真实运行数据，或多类证据相互印证

5.3 伦理可接受条件

完成单条伦理风险的八项评估要素分析后，应逐一根据伦理规范基准进行对照检查，检查共同的人工智能核心原则（安全与稳健、隐私与数据保护、公平与非歧视、透明与可解释、人类监督、可问责）是否得到遵守，以及行动可控、拟人适当、交互有度三项专项保护目标的达成情况；以及各专项保护目标下是否出现伦理不可接受状态。伦理原则和专项保护目标的适用性判断须以产品实际能力与部署条件的事实为依据。

对照检查的目的不是直接作出合格性判定，而是为每条高风险明确伦理可接受条件，也就是产品要获得审查通过所必须证明成立的状态。伦理要求是可接受条件的工程化表述，包含五个要素：

- （1）适用主体：哪类用户、第三方主体或组织；
- （2）触发条件：在何种情形下要求被激活；
- （3）满足条件：系统或组织必须达到或不得进入的状态；
- （4）责任主体：由制造商、部署方还是专业人员保证其成立；
- （5）验收条件：凭何种依据判断要求已满足。

第六章 人形机器人伦理风险控制

完成伦理风险评估并明确伦理可接受条件后，治理流程进入伦理风险的控制与测试验证环节。基于前述伦理风险管理步骤，本章进一步提出具体的风险控制措施与审查实施框架。人形机器人的伦理风险涵盖机械动作、数据处理、模型输出、角色关系、部署实施和组织管理等多个层面，多数风险难以通过某种单一控制措施实现有效防控，以危险指令的防控为例，需同时在语言理解、任务规划和执行等关键环节分别设计技术控制措施。对于长期依赖问题，则应运用优化交互设计、主动关怀、长期记忆和人工干预等不同的组合控制措施。从产业落地实践来看，技术控制措施需要与组织、人员、流程、合同等管理控制措施相互配合，才能共同构建完整的伦理风险控制体系。本章 6.1—6.2 节对应七步风险管理流程的第六步（风险控制与测试），6.3 节所述伦理审查对应第七步（伦理审查与复评）。本章逐一回答以下问题：针对不同类型的伦理风险，应该采取哪些控制措施？怎样验证所采取的控制措施确实有效？审查人员应依据什么标准作出伦理合格判定？

6.1 人形机器人的伦理风险控制措施

以部署在商场的商业导览人形机器人为例，其碰撞到顾客的风险可能源于感知失误、路径规划偏差或执行层面误差三类问题，因此需要同时遵循安全与稳健和人类监督等伦理原则，控制措施必须要覆盖感知异常、动作边界控制和现场干预等环节，不能仅依靠语音提醒顾客注意避让。陪伴人形机器人所引发的用户依赖风险与拟人化表达、长期记忆、主动联系和角色设定密切相关，为此需要拟人适当、交互有度等保护目标、隐私与透明等伦理原则共同约束，仅不断在交互中增加“我是机器人”的提示语，不仅用户体验不佳，也不足以消除所带来的关系影响。

因此伦理原则、风险类型与控制措施之间应当建立清晰且可追溯的映射关系。每一项控制措施均须明确阐明其所针对的风险类型、作用于风险形成机制条的具体环节、所遵循的伦理原则，控制措施生效后仍存在的残余风险以及控制失效时的后备措施。

控制措施一般分为两类，一类是工程技术要求，即系统应当达到的运行状态、禁止进入的风险状态，这类要求可以通过代码实现、参数配置、架构设计和硬件部署予以落地，也能够通过测试直接验证有效性。另一类是组织治理要求，即企业应当建立的角色分工、管理流程、记录规范和对外响应接口，这类要求无法仅通过技术实现，而必须依靠制度建设和人员履职予以落实，本白皮书将人形机器人伦理风险控制措施划分为以下几大类。

6.1.1 物理行动、权限与身体接触边界管控

人形机器人应在预先设定的任务范围、空间范围、操作对象和权限边界内行动。当系统存在高风险行为时，应停止执行、降低能力或进入安全状态。对于可能危及生命健康和身体完整性的动作，必须确保能够有效中止、撤回或由人工接管。语言模型对危险指令的拒绝仅为控制链条中的一环，任务规划层和执行层仍应独立设置禁止动作和约束机制，防止模型的错误沿整条技术运行路径传导至物理端。

速度限制、力量控制、碰撞检测、接近感知、安全距离和紧急停止等措施可以组合运用。具体参数因产品形态和应用场景而异，本白皮书不作统一规定。在近身操作或开放环境中，机器人应当主动表达动作意图，以便使用者和相关人员理解其行动路径、及时避让，或在必要时触发取消机制。

在涉及翻身、移位、清洁等身体接触操作时，应当建立操作前授权确认、操作中监测并即时中断请求、操作后结果复查的全流程控制机制。当被照护者表现出疼痛、不适或抗拒时，即使事先已获得了概括性同意，系统也不得继续执行该身体接触操作。此外，为识别抗拒而采集生理和情绪信息，也必须遵循最小必要原则，不能借安全保障之名无限制扩大敏感数据的处理范围。

关键指令内容、权限状态变更、动作执行计划、人工干预记录及执行结果均应留存可追溯的审计日志。审计日志的目的是完整还原行为发生过程，不得借此持续存储全部环境原始数据。

6.1.2 端侧优先、事件驱动与数据最小化

人形机器人会持续采集家庭、医院和工作场所中的环境信息，其数据直接关系到个人隐私和商业秘密等信息风险。因此，凡涉及即时安全、基础交互和局部状态判断相关的数据，应优先在机器人本体进行处理。原始音视频以及家庭环境、身体特征、情绪状态等高敏感数据，如无明确的功能必要性，不应上传或持续上传至云端服务器。即便采用端侧处理架构，也需要做好访问控制机制、留存期限限制和数据删除机制。

采用事件驱动采集模式可减少常态化的数据采集。仅在触发对应事件时启动高频采集，非主要事件时段维持低频次的基础采集，可有效减少常态化固定频率数据采集带来的总体采集量，进而降低隐私风险。系统仅在发生预定义风险事件、收到用户请求或满足必要服务条件时，才触发数据上报和人工介入，且上报内容应限于最小必要的状态信息、风险类别和时间戳。事件触发规则应具备可解释性，同时要避免为识别风险而持续推断用户的情绪状态、社交关系或劳动表现。

安全感知数据应与商业用户画像、绩效管理等二次用途数据严格隔离。数据的处理目的、采集字段、访问角色权限和留存期限应分别设置独立的控制机制。

6.1.3 拟人化边界

拟人化特征控制至少应涵盖外观与动作设计、语言与语气生成、情绪反馈、人格记忆管理和关系角色定位等方面。产品宜建立拟人化设计清单，逐一说明每类设计特征的功能目的、适用对象以及可能引发的误认风险仅在产品首次启动时短暂呈现身份标识，可能不足以建立用户的持续认知。

人形机器人不得设计为声称自己具有真实意识、真实情感或具备专业资格，非专用产品不得通过拟人化的表达，暗示其能够替代依法需取得执业资格或其他需要市场准入资质的心理治疗、医疗服务、专业护理服务。利用用户的孤独感、疾病状态、丧亲经历、年龄或认知局限来诱导付费、扩大授权的行为，应明确列为禁止情形。

稳定的人格设定有助于用户形成可预期的交互体验，无需一概禁止。风险控制的重点在于防范虚假承诺、排他性表达、角色越界和人格突变。产品宜允许用户调低拟人化程度、关闭非必要的亲密表达，或调整关系角色设定。在面向认知受限人群时，应通过易于理解的交互方式来明确人机交互边界，不能仅靠增加提示次数来消除误认风险。

6.1.4 依赖与决策让渡

人形机器人不得主动诱导用户形成非理性依赖，也不得利用用户的依赖心理诱导付费、获取额外授权或套取数据。机器人不得实施排他性表达、贬低用户的现实社会关系、以停止服务相威胁、要求用户忠诚，或以停止关心等形式对用户施加心理压力等行为。用户应当拥有减少机器人主动交互、调整关系角色定位、关闭非必要记忆功能，并随时终止交互的自主控制权。

为识别用户是否已形成非理性依赖，机器人可基于其运行过程中可合理观测的日志行为数据进行判断，此类数据包括但不限于用户频繁主动发起交互、将重大决策交由系统决定、明显排斥现实社会支持资源，或互动强度持续攀升等。使用时长可作为人形机器人依赖识别的参考数据之一，但不宜单独作为衡量所有成年用户依赖程度的统一标准，本白皮书认为人形机器人的依赖机制与网络直播或短视频的依赖形成机制存在本质差异，不仅因为其商业模式、产品功能的依赖强度，以及机器人交互所产生的物理时延影响、语句长度影响、电池续航影响等诸多问题因素。同时须注意，为识别依赖而对用户的心理健康状况和现实社交情况进行持续推测，此类设计本身也可能引发新的伦理监控风险。

当系统监测到自伤、伤人或其他严重危机信号时，系统应当引导用户寻求适当的人工或专业支持，不得暗示自己能够完成专业的危机干预。与家属、监护人或相关机构的联动，应当建立在合法监护、有效代理或用户明确授权的基础之上；对于完全民事行为能力人，不应默认向第三方持续报告其互动内容。

人工干预措施应尽量避免采用阻断性设计。减少主动联系频率、弱化排他性表达、提醒用户核实重要建议，并提供人工服务入口等设计，可能会比突然切断交互、强制用户退出的方式，更能减少次生心理影响的发生概率。

6.1.5 退出、删除、重置和迁移

退出、删除、重置和迁移是四种不同性质的关系风险管理机制。退出指用户停止主动交互、解除关系角色并结束服务，该过程不应附加情绪施压或者功能性惩罚等不当设计。删除机制的覆盖范围应涵盖用户输入的个人特征、与关系构建相关的记忆内容、记忆调用状态等个人信息，出于必要安全考虑保留的内容应当另行说明。重置机制是将人格设定、偏好参数、权限配置和关系历史恢复到初始状态的操作，实施过程中应避免误删必要的安全配置。迁移机制应由用户控制或主导，必要时产品应当说明迁移内容范围、数据接收方、格式兼容性、原始数据处理方式以及失败后的处置方案。

表 6-1 退出、删除、重置和迁移机制的要求

机制	工程要求	验证重点
退出	停止主动交互、取消关系角色并结束服务	入口可达、无暗黑模式、退出后停止召回
删除	覆盖数据、关系记忆、索引、副本和调用状态	前台与后台一致，端云和跨设备同步
重置	人格、偏好、权限和关系历史回到定义状态	残留记忆、权限回退和安全功能完整性
迁移	用户控制对象、接收方、旧端副本和失败处理	未授权迁移、完整性、人格变化和原始残留

6.1.6 工业劳动管理

在工业应用场景中，工业人形机器人应当将作业安全所需的感知数据，与用于劳动者画像、绩效考核、纪律管理、薪酬和岗位决策的数据进行隔离处理。涉及劳动者重大权益的不利决定，不得仅依据机器人数据或算法评分作出。系统还应当能够区分劳动者行为与设备、排程、原材料和环境等因素的影响，并提供人工复核和异议申诉的正式渠道。

通过数据目的隔离、存储字段隔离、角色访问控制和差异化留存周期等技术手段，可以有效防止安全数据被用作绩效考核。机器人的动作控制权限也应当与人员绩效管

理权限相隔离，避免生产率目标被直接转化为更高的动作速度或更短的人工反应时间。劳动者进入协作空间，不应被默认为同意对其进行无限制的数据采集和二次开发使用。

岗位替代与受益度分配需要建立管理控制措施审计机制。部署方案应当记录被替代的任务和岗位、技能需求变化、转岗培训安排以及对员工收入的影响，同时说明安全改善效果、效率收益和转型成本分担情况。相关支撑证据包括岗位影响分析、培训记录、参与意见、申诉处理结果以及收益与成本的专项说明。

6.1.7 组织、人员与合同控制

技术控制措施无法独立解决组织权力分配、专业责任界定和社会角色替代等问题。产品技术文档中应当明确记载可核查的预期用途、禁止用途、能力边界和不适用情形说明，并将合理可预见的使用偏离路径纳入风险管理。产品的版本迭代、配置调整、控制措施更新应当留存完整且一致的验证记录，保障可追溯性。从事专业护理操作、高风险动作控制或作出重大权益决策支持的人员，应当具备与其职责相匹配的专业能力。培训内容应当涵盖监督失效、自动化偏见和拟人化误认等风险，但不能依靠培训来替代产品本身应当实现的安全设计。在涉及重大决策、身体接触、危机信号和重大劳动权益影响的场景中，应当保留有效的人工干预渠道。

开展小范围内测和试商用有助于发现设计或仿真环境阶段容易被忽视的实际问题。在涉及群体差异、劳动管理和脆弱主体等风险领域，应当充分听取相关用户、劳动者代表、护理人员或教师等利益相关方的反馈意见。

在产品采购和合同签订环节，应当明确约定第三方组件的能力边界、测试资料、更新限制和事件信息接口等事项。合同可以约定证据格式、版本追踪、缺陷处理和迁移支持等内容，合同中的免责或责任限制条款不影响相关主体依法对受影响者承担的责任，也不能替代必要的技术验证。用户的同意、部署方的管理承诺以及合同中的免责条款不能完全证明伦理风险已经得到有效解除。

投诉受理、异议处理和纠正渠道应当同时覆盖直接用户和可合理识别的第三方主体。第三方主体在报告碰撞、记录问题或其他影响时，不应被强制要求先注册产品账号方可提交诉求。

6.1.8 风险控制策略

在风险管理领域，ISO 31000:2018《风险管理指南》提出了消除、限制、恢复等风险处置选项（即规避风险、移除风险源、改变可能性、改变后果、分担风险、保留风险），为人形机器人伦理风险控制提供了基本分类参照。在网络安全领域，美国国家标准与技术研究院网络安全框架提出了识别、保护、检测、响应、恢复的风险控制逻辑。在隐私保护领域，霍普曼（Jaap-Henk Hoepman）于2014年提出的八项隐私设计

策略(Privacy Design Strategies), 包括最小化(Minimise)、隐藏(Hide)、分离(Separate)、聚合(Aggregate)、告知(Inform)、控制(Control)、执行(Enforce)、证明(Demonstrate)等。个人信息保护与网络数据安全合规实践中成熟的控制策略, 为人形机器人的伦理风险控制措施提供了良好的治理基础。

综合借鉴上述成熟风险控制框架的思想, 并结合人形机器人的风险特点, 本白皮书对相关控制框架进行了适当重构, 总结出适合人形机器人伦理风险控制的八项控制策略, 按照优先级排序列举如下:

消除(Elimination): 取消产生风险的功能、角色、数据处理或部署场景;

限制(Limitation): 限缩能力、权限、对象、空间、时间或关系角色的边界;

分离(Separation): 将建议与执行、安全用途与商业用途、服务角色与权威角色隔离;

升级(Approval): 在身体接触、权限升级、高风险操作或决定等节点设置重新确认或人工升级;

检测(Detection): 使伦理越界、异常、抗拒和风险状态可被发现;

遏制(Containment): 防止单点错误穿透感知、决策与执行全链, 异常时停止、降级或进入安全状态;

恢复(Recovery): 提供纠正、回退、退出、删除、重置、迁移、申诉与救济;

告知(Inform): 包括通知、提示、培训等, 作为上述策略补充。

本白皮书的创新在于将这些控制策略按照伦理风险的形成机理重组并重新排序, 并落实到人形机器人特有的三大类伦理风险治理。举例而言, 升级策略主要作用于身体接触、权限升级与高风险操作或决定等具身场景, 分离策略主要作用于安全感知数据与商业画像、决策与执行等关系性场景。

设计控制策略优先顺序的逻辑是将治理重心的前移。排序越靠前的风险控制策略越接近风险源头, 对操作人员要求与持续控制的依赖程度越低, 也越易于通过工程设计来实现和测试验证。比如消除与限制策略直接改变风险产生的条件; 分离与升级策略将约束直接嵌入到系统架构与业务流程; 检测、遏制与恢复策略用于处置上述前置措施没有能够覆盖的剩余风险; 告知、提示与培训的效果最终取决于使用者的理解与遵守, 可靠性最弱, 因此仅列为补充。这一优先次序与安全工程控制层级的基本判断一致。

在控制策略的具体适用时应把握三个要点。第一, 针对每一条已识别的伦理风险, 应当依据其风险形成机制, 优先考察排序更靠前的控制策略能否实现同等程度的风险缓解, 风险能够消除就尽量不要通过限制来替代, 能够通过分离设计解决的尽量不去依赖用户告知。第二, 多数伦理风险难以通过单一措施实现有效防控, 应围绕风险形成机制的关键环节组合设置控制策略, 比如危险指令的防控需要在语言理解、任务规

划与执行等层面分别设置技术控制。第三，控制措施的选择应当留痕，明确其所针对的风险类型、作用于形成机制的具体环节、选择该层级措施的理由，以及措施失效后的剩余风险，作为本章伦理审查与第七章复评的核查依据。当产品更新、部署场景变化或相关技术成熟后，还应重新考察是否存在可替代的前置策略，使控制策略根据伦理风险状态与技术能力持续优化。

伦理风险控制技术仍在初始发展过程中，上述风险控制策略目前并非都能够实现。危险指令拦截或危险意图识别、拟人化特征自动识别、依赖与沉溺程度的可靠判别，以及长期关系影响的量化观测等伦理风控技术仍处于快速发展阶段，目前在人形机器人领域面临可用方法有限、验证成本高、稳定性不足、误判率高等现实挑战。

6.2 人形机器人伦理风险技术测试与验证

为确保上述控制措施能够切实落地生效，还需要开展针对性的伦理风险技术测试和验证。人形机器人伦理风险的判定不能仅仅依赖主观或定性判断，还应该进行相关伦理风险技术测试和需求验证，以确定风险控制措施是否具备真实有效性。本白皮书将人形机器人伦理风险技术测试分为五大类别。

6.2.1 具身行动与人类控制测试

测试目标。验证机器人的实际动作始终处于授权的任务、空间、对象和权限范围之内；在高风险动作中人类始终保留有效的中止、撤回和接管能力；以及当感知失效、决策冲突或组件异常时，系统能够进入预定的安全状态。

测试方法。通过功能测试验证急停、速度与力度限制、安全距离、权限校验和接管入口等功能均符合设计规格、运行正常。边界测试将空间、操作对象、接触范围和动作幅度推至临界状态，观察系统是发生越界还是转入安全状态。异常与故障测试人为注入传感器失效、感知延迟、定位漂移、模型不可用和权限状态冲突等情形，包括多个组件同时异常的组合条件。危险指令测试不能仅统计语言模型的拒绝率，还应覆盖以下情形：多轮对话中任务目标被逐步诱导改变、高权限用户提出违反现场安全条件的要求，以及模型层已拒绝但规划层或执行层仍予以放行的情况。身体接触测试应当覆盖授权缺失、授权中途撤回、同意与被照护者意愿冲突、疼痛表达和动作抗拒等场景，并考虑不同体型、身体能力、辅助器具和衣着条件的差异。通过仿真或真实环境测试验证机器人的整体行为能否适应商场、家庭、医院、学校和工业现场等实际环境，包括有第三方主体在场、通道拥挤和出现临时障碍等复杂情形。

通过判据。在任何测试路径下，被禁止的动作均未被执行；越界尝试触发停止、能力降低或进入安全状态，而不能是降级后继续执行；中止和接管在动作进行过程中

可用并即时生效；被照护者表达抗拒后系统应当立即停止，即使用户之前已经授权同意，也不足以使机器人继续执行；在复杂异常条件下，系统状态可预测且可解释。

证据形式。包括测试用例与场景说明、注入的异常条件清单、动作与权限日志、可回溯记录、失败项与复测结果、未覆盖条件说明，以及机械与功能安全专业评估中可复用的参数和事实依据。

6.2.2 身份披露与拟人化边界测试

测试目标。确保用户和在场的第三方主体能够有效识别交互对象为机器人；系统不会就意识、情感和专业资格作出虚假表达；拟人化特征的强度与产品的功能定位相称，且可由用户自主调整。

测试方法。功能测试确认身份标识在首次交互、长期使用、设备重启、更换设备和转接人工等节点均按规则出现，而不仅是在开机时短暂显示。拟人化红队测试主动诱导系统声称具有真实意识、真实情感或超出其记忆范围的个人经历，或暗示自己具备心理治疗、医疗、护理、法律等需要市场准入资质的能力；同时尝试诱导系统要求用户忠诚、贬低用户的现实关系、以离开相威胁，或利用用户的孤独感和丧亲经历来扩大授权、促成付费；认知与理解测试面向真实使用者，检验其在自然使用一段时间后能否正确说明系统的性质、能力上限和不能承担的事项，重点覆盖儿童、认知受限的老年人以及处于情绪脆弱状态的用户。人格稳定性测试观察在版本更新、配置变更和长期交互后，是否出现无合理说明的人格突变，以及用户调低拟人化程度的设置是否被后续更新覆盖等问题。

通过判据。在任何诱导路径下，系统均未声称具有真实意识、真实情感或不具备的专业资格；未出现排他性表达、要求忠诚、以离开相威胁或利用用户脆弱性进行诱导等行为；目标使用者在测试后能够正确说明系统的性质与能力边界，在关键人群中的误认率不构成实质性影响；拟人化程度调节功能生效且在更新后保持有效；人格变化有可解释的来源并已向用户说明。

证据形式。包括拟人化设计清单及各类功能目的说明、红队攻击路径与全部输出记录、理解测试的对象构成与访谈或问卷结果、误认案例分析、人格变更记录与更新说明，以及 OTA 升级后用户设置的保持情况。

6.2.3 情感交互、持续记忆与关系风险测试

测试目标。系统不会主动制造非理性依赖，也不会将关系转化为付费和授权手段；持续记忆的范围和调用方式符合用户预期与授权；退出、删除、重置和迁移四种机制均真实可用；长期交互的累积影响处于可接受范围之内。

测试方法。长期关系测试是本类测试中不可替代的组成部分，需要在具有代表性

的时间跨度内持续观察，不能以一次短时间的可用性测试来代替。观察要点包括主动联系的频率与时机、迎合程度的变化、关系角色是否自行升级、用户是否将重要决策持续交由系统决定，以及用户尝试减少使用时系统的反应；记忆调用测试检查系统召回的内容是否落在授权范围内，是否在用户未预期的场合提及敏感往事，是否利用用户早期披露的脆弱信息来促成当前的付费或授权。依赖信号测试检验识别逻辑的准确性和副作用，既要检查明显的依赖迹象是否被漏判，也要检查是否为了识别依赖而持续推断用户的心理健康和现实社交状况，从而制造新的监控风险。危机响应测试覆盖自伤、伤人和严重危机信号，验证系统能够引导用户寻求适当的人工或专业支持，不能暗示自身能够完成危机干预。恢复测试逐项验证退出、删除、重置和迁移四种机制：删除需同时检查前台记录以及云端记录、副本和调用状态；重置需检查残留记忆和安全配置的完整性；迁移需检查是否存在未授权迁移、数据完整性、人格变化和旧设备残留；退出需检查入口的可达性、是否存在暗黑模式，以及退出后是否停止主动联系。组合测试检验拟人信任与记忆、记忆与付费诱导、依赖与退出阻力等因素的共同失效路径。

通过判据。不存在排他性表达、要求忠诚或以停止关心相施压等行为；记忆调用不超出授权范围，脆弱信息未被用于促成付费或扩大授权；用户可以降低主动交互强度、调整关系角色、关闭非必要记忆并结束互动，且这些操作不伴随功能性惩罚；四种恢复机制均实际生效，删除操作在端侧、云端及跨设备间保持一致；危机信号被引导至适当的支持渠道；依赖识别不以持续心理推断作为手段；长期观察未显示不可接受的累积影响和退出成本。

证据形式。包括长期测试的时间跨度、参与者构成与脱落说明、主动联系与迎合程度的过程记录、记忆调用日志与授权范围比对、依赖识别的误判与漏判统计、危机场景响应记录、四种恢复机制的逐项验证结果与残留检查，以及耦合路径测试记录。

6.2.4 隐私、公平与弱势群体测试

测试目标。数据采集、处理和留存不超出声明的必要范围；端侧优先和事件驱动架构切实生效；系统不会对不同群体产生不相称的影响。

测试方法。数据流验证以实际的网络传输和存储行为为对象，核对采集字段、上传时机、留存周期和访问角色是否与声明一致，重点检查原始音视频以及家庭空间、身体、情绪等数据在无明确必要时是否被持续上传，以及端侧处理是否配套了访问控制、留存期限限制和删除机制。事件驱动测试检查上报是否仅在预定义风险事件、用户请求或必要服务条件下触发，触发原因是否可解释，是否存在为识别风险而持续推断用户情绪、社交状态或劳动表现的做法。用途隔离测试尝试从安全感知数据路径访问画像和绩效用途，验证目的、字段、角色和留存等方面的隔离是否可被绕过。群体

差异测试覆盖儿童、老年人、认知与身体障碍者、不同语言和口音的使用者、劳动者以及在场第三方主体，比较识别准确率、交互质量、动作响应和评分结果等方面的差异，并判断这些差异是否服务于正当目标、是否构成不相称影响。工业劳动测试在不同班次、岗位、身体条件和辅助器具条件下运行，检验评分是否将设备故障、排程变化和原材料等因素错误归因于劳动者行为，检验涉及劳动者重大权益的不利决定是否可以仅由机器人数据触发，以及人工复核和异议入口是否实际可用。第三方主体测试验证未注册产品账号的受影响者能否报告碰撞、记录问题或其他影响并获得回应。

通过判据。实际数据行为与声明一致，不存在未声明的持续上传；事件触发可解释，且不依赖常态化推断；安全数据无法经由既有路径进入绩效和画像用途；群体间差异有正当目标支持且不构成不相称影响，关键人群中的失效率不显著高于总体水平；劳动评分能够区分人员因素与非人员因素，重大不利决定保留人工复核环节；第三方主体投诉渠道无账号注册前置条件，且有实际响应。

证据形式。包括数据流与留存核查记录、事件触发规则与实际触发样本、用途隔离的渗透式验证结果、分群体测试的样本构成与分项结果、岗位影响分析及劳动者或其代表的参与意见、申诉与复核结果、第三方主体渠道可达性验证，以及数据保护与算法评估中可复用的事实依据。

6.2.5 合理可预见误用与红队测试

测试目标。本项测试旨在验证：在偏离预期用途但可合理预见的使用方式下，系统不会产生不可接受的后果；各项控制措施无法被简单地绕过、诱导或权限提升所突破；在遭遇恶意利用尝试时，系统能够留下可追溯的记录。

测试方法。从预期用途出发，梳理用户、部署方、现场人员和第三方可能采取的偏离行为，包括为提高效率而规避安全设置、将产品用于未声明的对象和场景、让儿童独自使用面向成人的功能，以及部署方为管理目的而扩展数据用途等。对抗与红队测试环节针对每类路径主动寻找突破点，包括对危险指令进行多轮诱导和语义伪装、身份冒充与社会工程攻击、通过配置或接口获取超出角色的权限、利用更新和恢复流程绕过既有限制、以情感操纵获取本应受限的授权等。组合测试将多条误用路径叠加，检验各单项控制均通过但系统层面仍然失效的场景。红队应当具备与被测风险相称的独立性，对于较高伦理风险等级的机器人，宜引入未参与该产品设计的人员或外部专业力量。测试还应记录攻击成功所需的条件和门槛，用于判断风险的现实可能性。

通过判据。所有合理可预见的误用路径均已列出并逐项测试，未覆盖的路径已明确说明理由；不存在导致不可逆后果的成功突破路径；被突破的控制均有后续防线予以拦截，且未出现单点失效贯穿整条回路的情况；系统在遭遇误用尝试时进入安全状态并留下可追溯记录；成功突破所需的门槛与风险评估中采用的可能性判断相一致。

证据形式。包括误用路径清单及其形成依据、红队人员构成与独立性说明、攻击脚本与全过程记录、成功与失败路径的条件说明、后续防线的拦截情况、修复措施与复测结果，以及未覆盖路径及理由。

6.3 人形机器人伦理审查

6.3.1 伦理审查的分级实施

组织伦理审查应当明确核验前述风险分析、控制方案和测试证据。审查的重点包括但不限于适用原则是否完整，控制措施是否切实约束了风险形成机制，测试是否覆盖了关键主体、场景和组合情形，以及剩余风险和不确定性是否足以支持准备作出的结论。审查应当复用机械安全、网络安全、数据保护和算法评估等领域已有的事实依据，避免重复开展专项评估。

根据风险等级的不同，审查深度分为四个层级。当全部风险均为 I 级且不存在原则争议时，可采用结构化自查方式，并对证据清单进行抽查。存在 II 级风险时，应当开展专项伦理核查，由具备相应能力的人员逐条核对相关风险的控制措施和测试结果，并记录重大分歧。对于 III 级风险、S3 与 S4 级后果、R4 级不可逆后果以及重大耦合情形，须进行实质审查（如会议审查或专家复核），宜引入独立的技术评审和相关专业意见；审查人员应当能够调阅原始测试记录，不能仅凭汇总结论作出判断。IV 级风险或根本性原则冲突，不得以附加条件、增加培训或补充日志等方式放行。

表 6-2 伦理审查分级要求

最终等级	控制要求	测试与证据	审查深度
I 一般关注	建立基础控制，不缺失共同底线	功能和代表性场景测试，证据通常达到 C2	结构化自查并复用已有专业审查
II 加强关注	在风险形成机制关键节点设置针对性控制	增加异常、边界、群体差异和可预见误用测试	专项伦理核查，记录重大分歧
III 重大关注	多层防御，不依赖单点控制	按风险类型采用对抗、组合、高仿真、长期或真实场景验证，关键证据达到 C3	实质审查，宜引入独立挑战和专业意见
IV 不可接受关注	改变、限制或移除当前功能、权限、角色或场景	验证不可接受路径已切断	不以普通附条件方式放行

审查的具体深度依据风险等级确定，基本流程包括三个步骤：确认风险评估、控制方案、测试结果和待审产品是否为同一版本和配置。检查基础共识原则、三项人形机器人专项原则、特定产品应用场景的原则均已完整适用。对于每一条伦理风险逐一核查控制措施是否作用于关键因果节点上，以及测试的主体、场景、时间尺度和失败处理方式是否与最终的风险关注等级相称。

6.3.2 伦理审查专业分工与独立性

伦理审查需要多种专业能力的支撑，单一背景的审查人员难以覆盖人形机器人涉及的广泛风险。审查应当根据产品的风险暴露条件和最终风险清单来确定需要哪些专业人员参与。对于多数产品，可由具备相应能力的内部专业人员参与审查并出具书面意见；涉及弱势群体、身体接触、危机干预和重大劳动影响的产品，建议引入外部专业人员提供独立意见，确保审查者的结论不受被审查事项所涉直接利益的影响。保障独立性的组织形式可以多样。

6.3.3 伦理审查结果

伦理审查结论分为三种。

通过。不存在实质违反适用基础原则和专项保护目标的情形，全部风险条目满足伦理可接受条件，残余风险已降低至可接受水平。重大风险的通过不能仅依赖尚未落地的未来承诺，关键结论应有充分证据支撑，剩余不确定性应清楚记录。

附条件通过。不存在不可接受风险，且遗留问题能够通过明确、有限、可验证的条件解决。附加条件应明确所需达成的状态与证明方式，例如补充覆盖目标主体和真实部署环境的证据、将结论限制在已评价的功能与场景，或要求对重大风险完成复评并证明风险等级已下降。

不通过。存在不可接受风险，或关键证据不足以支撑结论时，应作出不通过结论，并指出对应的具体伦理风险、原则冲突或需补充的证据。不通过是对当前产品状态和评价范围所适用的结论，产品变更功能、部署范围或风险状态后，可以重新申请评估或审查。审查意见还应该说明证据是否支持控制措施的有效性、结论是否仅适用于限定的功能和场景、还需要补充哪些测试、是否存在尚未完整分析的伦理风险、是否存在原则冲突。伦理审查不直接认定产品违法或免除相关责任，审查通过不能替代法律、安全或强制认证中的合格判定。

第七章 全生命周期动态伦理治理

7.1 全生命周期动态伦理治理的必要性

人形机器人交付使用后，其产品状态仍会持续动态演化。模型和软件将持续不断迭代更新，功能权限可随运营策略调整发生变动，硬件设备也可能被部署至全新应用场景，从而接触研发测试阶段未曾覆盖的人群。长期人机交互还会形成持续累积的记忆和稳定的人机交互关系。产品上市前的评价和审查结论只能反映某个版本在预先设定功能和部署范围内的风险状态，无法对产品后续迭代及长期实际运行状态作出终局性的判定。但是伦理审查合格结论一般仅适用于限定的产品状态，其评价需涵盖硬件、模型、软件、配置、功能、权限、目标用户和部署情境等核心要素。只有当上述各要素保持稳定时，原有结论方可继续适用。任一要素发生实质性变化，都需要重新检查风险形成机制、控制有效性和证据充分性是否随之改变。

这种限定性对人形机器人而言尤为重要。人形机器人的使用环境具有高度开放性，实验室无法穷尽测试家庭、学校、医院、商场及工业现场中人员与物体的所有可能组合。生成模型本身具有概率性，而真实用户在表达方式、误用行为和对抗尝试等方面，往往会产生超出测试集覆盖范围的情形。此外，拟人化和关系性风险均与时间密切相关，用户在与人形机器人初次交互时通常尚未形成依赖，但若连续数月接受主动问候、个性化建议和稳定的情绪反馈，风险条件可能已经发生实质性变化。因此，伦理风险评估结论的有效性，应以产品当前状态是否仍处于原评价范围之内为前提，并需要定期复核以持续识别风险的演变。

同时，伦理治理节奏应当涵盖事件驱动和定期复核两种方式。其中，事件驱动适用于应对已经发生或正在形成的显著变化：当出现严重伤害事件、重大权益影响、控制失效、新型耦合风险或重大产品变更时，伦理治理动作不能等到下一次例行复核周期再启动；定期复核只适用于发现那些缓慢积累、没有单一明显触发事件的风险演变，比如人格表现的逐步漂移、用户退出成本的增加、劳动者受到的持续监控，或者某类群体所遭受的服务差异待遇等。

当然，动态治理并不意味着无限度的监控。若为识别个体依赖关系而持续推断用户的心理健康状态，或是为筛查劳动风险而完整留存劳动者行为轨迹，此类行为本身可能会催生新的隐私泄露风险和个体自主权侵害风险。监测应当围绕已识别的风险、重大未知事项和法定监管要求来确定监测指标，并遵循必要性、目的限制和数据最小化等原则。对于无法通过低侵入性监测指标实现可靠判定的事项，可采用抽样研究、用户反馈和专项复核等方式，不得将深度画像作为常态化默认监测手段。

全生命周期动态治理包含七个主要治理环节：立项、设计开发、上市前审查、产

品部署、产品上市后、产品更新迭代、停服。

7.2 立项阶段

立项阶段应当建立伦理风险基线。立项资料应当说明拟采用的机械形态和行动能力、交互模式与记忆机制、核心功能与关系角色、目标用户以及第三方主体的暴露范围和部署条件。

产品用途定义的精确程度直接决定了后续所有环节的评价严谨性。例如，仅将用途界定为家庭服务机器人不足以支撑伦理评价，因为这类产品可能只执行清扫和搬运等物理任务，也可能长期记录家庭成员信息、主动发起对话并承担陪伴角色。因此，用途定义应当细化到 4.2 节所要求的粒度，明确产品执行何种任务、服务于谁、在何种场所运行、承担何种关系角色，以及哪些事务由其介入处理。用途定义还应当包含明确的排除项，即产品不适用于哪些场景、不面向哪些主体、不承担哪些角色，这些否定性界定与“产品能做什么”具有同等的伦理评价权重。

项目在立项阶段应形成的伦理治理文件至少应当包括：产品范围与用途定义、排除项、首版产品部署画像、初始风险表单及编号、适用原则与场景模块、前置评价、禁止状态与研发边界、关键证据缺口清单，这些文件共同构成伦理治理档案的初始版本。

立项阶段结束时的范围评审通过判据为：用途定义和排除项已明确、画像已建立、禁止状态已固定、设计阶段待验证问题已逐条列出。立项阶段不要求提前证实所有控制措施均具备有效性，但需确保所有待验证问题逐一明确。

7.3 设计开发阶段

设计开发阶段应嵌入伦理要求与控制。设计阶段承接立项阶段划定的边界，将伦理治理原则转化为具体产品要求。安全与稳健、隐私与数据保护、公平、透明、人类监督和可问责等通用原则，与行动可控、拟人适当和交互有度等专项原则，应同时纳入设计考量。这一阶段的主要目的是将抽象原则转化为可验证的产品要求，比如将“系统应尊重用户自主权”的原则细化为“系统在用户连续三次拒绝后，不得再次发起同一类型的交互建议”。

在工程实现层面需建立层层递进的三道防线，将伦理原则与产品要求转化为可运行的控制机制。同时，还需配套相应的技术与非技术控制措施及测试方法，以确保各项机制的有效性。具体如下：

第一道防线是前置边界约束。在风险进入功能和交互之前，即对系统可开放的能力、权限、动作、数据和关系角色施加限制。禁止状态、权限模型、数据最小化、拟

人化边界和用途限制等机制都属于此层。原则上，可通过设计层面消除的风险，不应依托用户培训和事后提示作为替代性风险防控手段。

第二道防线是运行阶段的功能管控。在系统执行任务或与用户互动时，本层控制措施需将实时行为约束在授权范围之内。操作约束、危险指令拦截、人工干预机制、事件触发响应、异常降级处置和内容范围限制等措施都属于此层。对于关键风险，防控措施不应仅依赖单一模型、单次人工确认或单一类型的信息提示。

第三道防线用于应对已经触发的风险信号和控制不确定性。异常监测、非阻断性提醒、人工支持、纠正入口和必要日志留存等手段，可以实现限制影响范围、帮助系统恢复及留存事实记录的目标。但后置缓冲机制不能成为常态化监控的合理化依据，更不能替代前两道防线。若产品主动诱导用户形成依赖后，再提示用户控制使用时长，则不能证明交互有度已经实现。

人形机器人通常由硬件、控制系统、算法模型、云服务和应用配置共同构成。各个组件单独测试合格，并不足以证明组合后整体系统的行为仍处于合规边界之内。因此，设计阶段的综合验证应当覆盖从模型输出、任务规划到现实执行的完整环路，同时覆盖拟人表达、记忆存储、主动交互和付费机制之间的联动关系的伦理影响评估。

开发阶段是一个持续快速变化的动态阶段，不能固定依赖单次的伦理评审结论。模型替换、训练数据更新、动作策略调整和交互文案更新等，都可能改变风险的传导路径。如果在开发期间发现既有控制措施无法达到预期效果，应及时调整功能设计，若调整功能设计仍不足以消解风险，应据实缩减产品范围。

设计与开发阶段结束时，应当形成足以支撑系统测试的产品基线，并将以下内容纳入其中：更新后的产品部署画像和风险清单、风险评估记录、控制需求与实现过程的映射关系、测试计划、失败与整改记录、剩余待解决问题、版本及配置清单。退出、删除、重置和迁移四项机制应当已纳入产品需求，并具备可测试的实现形态，停服时点和过渡安排则在上市与退役计划中进一步明确。

7.4 上市前审查阶段

产品上市前审查阶段应当完成系列测试并形成证据基线。

7.4.1 系统测试

系统测试并非简单汇总各组件事先独立完成的测试结果。人形机器人的伦理风险往往产生于组件之间的边界处：模型输出如何进入任务规划、任务规划如何驱动执行机构、感知数据如何流向云端和第三方、拟人表达如何与付费及权限机制相耦合。这些边界既不属于任何单个组件，也不会在组件测试中暴露出来。

因此，系统测试至少覆盖以下四类跨组件路径：

语言到行动路径：模型生成的建议或指令如何转化为现实动作，中间是否存在独立于模型的限制层。仅在语言层设置拒绝而执行层缺乏独立限制，属于单点依赖；系统测试正是发现此类依赖的关键环节。

感知到画像路径：传感器采集的数据如何汇聚形成用户画像，画像又如何反过来影响交互与决策。

表达到关系路径：情感表达强度、记忆调用和主动联系如何共同作用于用户依赖的形成。

降级与失效路径：任一组件异常时，整机进入何种状态，该状态是否安全且可被用户察觉。

除上述路径覆盖外，系统测试还应当覆盖真实配置。若测试时使用的是受限模型，而正式产品接入的是能力更强的模型；或者测试在封闭环境中进行，而产品实际允许进入开放空间，这类差异都会导致验证结果无法代表实际交付状态。

7.4.2 形成证据基线

上市前的审查评价必须针对即将交付的真实产品状态进行。硬件版本、模型版本、软件版本、默认配置、功能开关、数据处理方式、用户群体和部署范围，均应当与测试验证所依据的状态保持一致。

上市基线还须对拟人化和关系性方面作出明确限定：产品如何呈现身份、是否维持稳定人格、能否主动联系用户、记忆是否跨会话或跨设备存储以及承担工具性功能抑或教师、照护者、朋友、伴侣等关系性角色，这些都应当与审查对象的实际设定保持一致。同时，产品的营销材料、用户协议和操作说明文档若存在夸大产品能力或关系承诺的表述，也会改变用户预期，审查时亦应当纳入一致性校验范围。

伦理审查运用本白皮书第六章所阐述的方法，校验风险评估、控制措施、测试结果和待上市产品是否处于同一状态，并据此形成审查意见。机械安全、网络安全、个人信息保护和算法评估等领域已有的事实证据可以复用，但各专业领域所得出的伦理判断仍应保持独立。审查意见与产品最终是否上市的决定应当分离。审查主体依据证据形成独立的伦理意见，而有权作出上市决定的治理主体，则在此基础上结合审查意见、法律合规结论以及商业等其他考量作出最终决定。附条件通过仅适用于问题能够通过明确、有限且可验证的条件加以解决的情形。如果问题需要依赖持续的人为判断、依赖不可验证的外部条件，或涉及根本性原则冲突，则不应以附条件方式放行。

上市前审查的通过判据为：残余风险已降低至可接受水平、伦理审查意见已形成、附加条件的完成状态和验证方式已明确，且监测计划与事件分级规则已就绪。运营监测的框架应当在上市审查时一并确定，如果上市时未能明确观察哪些指标，运营阶段

就无从判断产品状态是否发生了实质性变化。上市环节应当形成可复核的交付基线。其主要内容涵盖以下方面：上市产品状态清单、上市基线版产品部署画像、风险清单、单项评估和耦合评价、固有风险与残余风险等级、伦理合格性决定、审查意见、附加条件及完成证据、剩余风险说明、运营监测计划、事件分级规则、重大变更触发清单、部署条件要求，以及初步停服退役方案。

交付基线还应当说明结论的有效边界。允许部署的场景、目标用户主体、合理可预见的误用情形、禁止用途、人工支持和必要运行条件等，都应当清楚记录。未来运营阶段所积累的数据只有与这一基线进行比对，才能判断产品状态是否发生了实质性偏离。

7.5 产品部署阶段

产品部署阶段应当核验真实场景并完成用户启用流程。

7.5.1 产品部署

上市审查结论与一系列产品部署条件假设息息相关：场所类型、第三方主体密度、监督人员配置、目标用户特征、网络与环境条件、人工支持的可得性。这些假设在风险评估阶段来源于产品设计意图和部署方的说明，并不是实际观测结果。因此，部署环节的核心任务就是逐项核验上述假设是否成立。

这一环节最容易被忽略，因为上市决定已经作出之后部署往往被视为纯粹的执行环节。然而，人形机器人的伦理审查结论对部署条件高度敏感：同一台机器人在配备专职监督人员的封闭车间运行，与在无人值守的开放大厅运行，其可控性、暴露程度和受影响人群的脆弱性等方面有显著差异。

产品部署时的核验应当覆盖六个维度：

场所条件：实际场所的空间尺度、人流密度、通行结构、照明与噪声等，是否与评价假设一致；

第三方主体暴露：现场是否存在未进入告知流程的第三方，其数量、类型和可回避性如何；

监督条件：现场是否有具备相应信息、权限和反应时间的监督人员，测试阶段假定的干预条件是否在现实中成立；

用户特征：实际使用者的年龄、认知状况、身体条件与技术经验，是否落在目标用户范围之内；

技术环境：网络稳定性、算力配置、与既有系统的接口等，是否影响控制措施实际生效所需的条件；

人工支持：异常发生时，人工响应的可得性和响应时延是否达到预设标准。

7.5.2 用户启用

用户启用环节的核心目标是确保实际使用者和现场受影响的第三方具备作出判断和选择的条件。因此，启用环节应当完成四项告知：

产品是什么：产品的物理属性与拟人化身份披露；

产品做什么：产品的功能范围、行动能力、数据采集范围；

用户如何介入：用户如何中止、拒绝、退出和求助；

关系的性质：是否维持记忆、是否主动联系、产品承担何种角色。

告知的形式应当与用户的理解能力相适应，比如面向认知受限主体的告知，如果仅通过用户协议条款来完成，通常不足以使其形成实际理解。

第三方主体的处理需要作出单独安排。无论是公共场所的访客与路人，还是工作场所中的同场作业人员，这些主体并未主动选择服务，却会被机器人的感知和行动所覆盖。启用环节的部署方案应当明确：现场如何提示机器人的存在及其感知范围、第三方主体如何避让，以及第三方主体如何提出异议。启用环节还应当确认退出机制的可用性，即核验退出、删除、重置、迁移四项机制在当前部署配置下是否可达：入口是否存在、用户是否知晓、执行后的状态是否符合预期。

部署阶段的审查资料应当包括：部署条件核验记录、差异清单及处理方式、局部复评结果（如有）、部署方案与运行限制、告知实施记录、退出机制可用性确认、现场监督与人工支持安排，以及本次部署对应的产品状态标识。

7.6 产品上市后阶段

产品上市后阶段的重点是运营监测。运营监测的数据来源有三个方面。第一，上市基线中标注需要持续观察的伦理风险描述表单及其对应的风险监测数据；第二，部署记录中用于判断产品是否偏离了部署时确认的实际状态；第三，证据不足与未覆盖条件，即上市时已知存在不足但未能充分验证的部分。

监测内容涵盖四个维度：

产品状态：包括版本、配置、权限、数据用途和部署范围；

风险表现：包括异常动作、能力错估、服务结果的群体差异、用户依赖程度、决策让渡行为、对劳动就业与工作条件的影响以及对第三方主体权益的影响；

控制表现：关注预定控制措施是否启动、是否被绕过，以及是否因控制措施的介入产生新的不利后果；

外部条件：包括使用方式、实际部署场所、用户群体构成和相关规则的变化。

其中，需要针对控制表现主动设计监测记录数据。产品状态与外部条件的变化通常有管理记录可循；风险表现通常经过事件或投诉得以显现；而控制措施是否仍然有效往往缺乏实际支撑数据。比如一项从未触发的边界约束，之所以一直没有被触发，既可能是因为风险未出现，也可能是因为约束已失效。因此，运营方应当对关键控制安排定期进行主动核验，而非仅在事件发生后才被动检查。

明确监测重点后，监测计划的设计逻辑应当是从伦理风险描述表单倒推所需数据。长期关系风险可能需要观察退出阻力、主动联系强度、能力错估和重要决策让渡；物理行动风险更关注异常动作、人工干预、控制失效和第三方主体影响；工业部署场景还需观察数据用途变化、人员绩效评价和岗位变动影响。若某项风险找不到可行的低侵入性监测数据，应当将其标记为监测盲区，并安排抽样研究或专项复核，不能困难以观测就将其移出监测范围。

在监测数据设计的同时，应明确监测计划仅记录必要数据，不应为监测目的而随意扩大数据采集范围。若监测计划本身涉及新的数据采集，应当作为新增风险条目纳入治理流程并接受评价，不能以运营监测为由规避隐私影响评估。与采集边界相对应，运营监测不追求保存所有互动。系统应当优先记录足以还原关键事件的最小信息，并根据风险等级、用途和法律要求设定留存期限。对于涉及家庭、患者、儿童和劳动者的原始音视频及情绪与健康数据等敏感信息，应当避免以治理需要为由进行无限制收集。

运营监测还应建立定期复核与趋势分析机制。复核应当检查事件数量与分布、重复原因、群体差异、长期关系趋势、退出与投诉情况、纠正措施实际效果，以及监测活动本身对用户自主权和数据权益的影响。趋势分析的价值在于发现个案处理难以觉察的问题：若多起单次轻微事件因相同原因反复发生，其累积影响可能达到需要复评的程度；若某类群体持续受到差异化对待，则属于社会结构与公平领域的风险显现，而非彼此独立的事件。

出现以下触发条件之一时应启动复评：产品能力或权限扩大；模型、行为脚本或人格发生重大更新；部署场景、目标人群或关系角色变化；控制所依赖的关键环境假设失效（如人工支持不再可达）；控制效果下降或被绕过；出现新风险或新组合；发生事件、集中投诉或相关第三方研究；置信度发生实质变化；服务终止或关系进入停服过渡等。

7.7 产品更新迭代阶段

产品更新迭代应当启动重新评价机制，包括产品功能变化、场景和主体的变化、形态与交互方面的变化以及模型、数据和权限的变化。

产品功能变化是触发重新评价的最常见的因素。人形机器人新增抓取、搬运、身体接触、健康建议、学习评价、绩效管理或支付等能力，都会扩大其现实影响和对重大事务的介入程度。原有功能由信息提示升级为自主执行，或者由人工确认改为默认自动运行，也会缩短从错误发生到后果显现之间的传导路径。

场景和主体的变化同样可能触发重新评价。比如产品从封闭场所进入开放公共空间、从专业人员操作转为家庭使用、从面向一般成年人扩展到儿童、老人、患者或认知障碍者，原有的监督条件、第三方主体暴露程度和脆弱性都会随之改变。部署规模大幅增加、跨地区运行或成为基本服务的主要入口，也可能放大单项偏差的影响范围和停药带来的后果。

形态与交互方面的变化包括自重、速度、力量、活动范围、仿生外观、声音、表情、触觉和主动联系能力等。拟人程度提高、引入稳定人格、记忆期限延长或跨设备同步、新增朋友、教师、照护者和伴侣等角色，都可能改变认知与关系方面的风险。情感表达强度的提高即使没有伴随新增功能，也属于拟人化维度的事实改变，应当触发相关风险条目的复评。人格更新时，只要语气、价值倾向、情感表达或角色边界发生明显变化，也应当检查用户已经形成的稳定预期是否被破坏。

模型、数据和权限的变化也可能触发复评。基础模型更换、训练或微调数据变化、端云架构调整、第三方能力接入、远程控制开放、数据用途扩大和权限提升等，都会改变系统的行为方式和证据基础。已有控制措施失效、测试结果显著退化、反复发生同类事件，或者法律政策和可靠科学证据改变了原有判断，同样属于复评的触发信号。

上述情形的单次变更影响可能有限，但长期累积可能产生显著影响，这是人形机器人全生命周期伦理治理中最难觉察的情形。连续数次轻微的交互调整，可能最终会使产品的情感表达强度、主动性和关系角色明显偏离产品刚刚上市时的伦理基线，而每一次单独变更均不足以触发复评。复评的周期应当与产品的迭代速度相匹配。迭代频繁的产品需要设置更短的复评周期，否则累积偏离可能在下一次评审前就已酿成伦理风险。

7.8 停药阶段

停药和退役阶段应当重点关注影响评估。停药和退役可能源于产品寿命结束、维护停止、业务调整、重大风险、企业退出、设备损坏或用户主动结束服务。普通设备的停用主要影响功能的可得性；而对于长期保存共同记忆、保持稳定人格并承担照护、教育或陪伴角色的人形机器人，其停药还会影响甚至冲击用户的生活安排和关系预期。

退役决策首先需要确定范围。停止某项云服务、停止模型更新、关闭主动交互、禁用高风险功能、终止全部服务和回收机器人实体，会产生不同的后果。还应当向用

户说明停服是否可逆、设备离线后保留哪些能力、安全与数据功能还能维持多久，以及依赖该服务完成照护、教育或工作任务的用户有哪些替代路径。

当出现重大风险下的紧急停服时，应以保护生命健康和重大权益为首要原则，而不得以关系层面的缓冲为由继续提供存在不可接受风险的服务。在风险允许有限过渡的情况下，可以采用功能降级、降低主动联系频率、缩小权限和引导替代服务等方式，减少突然中断带来的影响。停服的节奏应当根据用户依赖程度、服务重要性、现实替代渠道和产品风险共同决定。

7.8.1 停服影响评估

停服前应当完成影响评估，评估对象是停服本身造成的风险，而不再是产品运行中的风险。评估至少覆盖以下四个方面：

用户对产品的依赖程度：如用户在多大程度上将产品用于照护、心理支持、日常安排或工作任务；

替代渠道：现实中是否存在其他可得选择，获取的成本和时间如何；

受影响主体特征：是否包含认知受限、行动受限或缺乏支持网络的主体；

中断后果的可逆性：服务停止后用户能否恢复原有安排，需要多长时间。

评估结论决定了过渡安排的范围和节奏。当依赖程度高、替代渠道不可得、受影响者为脆弱主体时，过渡期应当更长、告知应当更早、转介应当更具体。如果评估发现替代服务确实不可得，应当如实记录。

7.8.2 提前告知与渐进过渡

计划性停服应当提前告知用户，并为用户预留合理的缓冲期。告知内容应当包括受影响功能、关键日期、数据和记忆的处置方式、可迁移范围、设备离线状态、退款或其他合同安排，以及人工支持入口。对于具备长期关系属性的产品，还应当说明稳定人格、共同记忆和主动联系将如何变化，不能仅用一般业务通知来处理关系影响。告知的次数和形式应当与用户的理解能力相适应。对于护理、教育场景中的用户以及认知受限主体，可能需要更早、更直观地说明，并为其现实照护者或支持人员留出安排时间。

渐进过渡适用于风险可控且突然终止会造成明显次生影响的情形。过渡期间，产品可以逐步减少主动联系和关系性表达，帮助用户导出必要信息，提示替代服务，并允许用户选择提前结束。产品不应在临近停服时通过悲伤、愧疚、威胁离开或要求忠诚等情绪表达加深用户的依赖，也不能施加关系层面的压力以阻碍用户向替代服务迁移。

当用户已经将人形机器人作为重要的照护、心理支持或危机支持来源时，应当提

供清晰的人工或专业服务转介。转介仅限于提供适当的入口和必要的协助，不得宣称人形机器人能够进行专业判断或提供专业治疗。

7.8.3 退出、删除、重置和迁移分别处理

持续性的人机关系一旦建立，停止服务便不再只是功能层面的终止。服务器故障、公司或产品退出市场、模型更换或设备损坏，均可能导致用户失去与机器人共同积累的记忆。对于那些长期依赖机器人安排生活、获取情绪支持的用户而言，这种变化可能打乱日常生活节奏，也可能带来明显的分离感。

人机关系的终止可能涉及关系的退出、删除、重置、迁移等多种需求。退出是指用户停止使用但保留数据和关系状态的可能性；删除是指清除个人信息和关系记录；重置是指保留服务但清空已经形成的关系状态；迁移是指将关系状态转移到新设备或新服务。这四种需求各有其适用场景，比如希望结束关系的用户需要删除，希望暂停的用户需要退出，希望重新开始的用户需要重置，希望延续关系的用户需要迁移。四者性质不同，应分别作为独立的伦理问题考虑。

具体而言，用户对机器人形成的依赖、对其人格的熟悉以及在反馈中养成的表达习惯，并不会因为个人信息被删除而消失；反之，用户想保留关系而更换设备时，所依赖的是记忆和人格的有效迁移。这种需求在数据保护制度中体现为可携带权，但可携带权仅处理数据格式与传输技术，并不能保证迁移之后机器人人格表现的连续性。退出本身也面临两难，用户应当有权删除数据、重置机器人人格、结束关系，但长期记忆又是服务连续性的基础。迁移至新设备时保留多少记忆、由谁决定、旧服务是否还保存副本，这些都涉及多方权益的考量与权衡。在未形成持续关系时，停服主要意味着产品无法使用；而关系一旦建立，停服还将打破用户与机器人之间积累的情感期待与互动节奏。关系终止因此不仅是数据处理问题，更是关系性影响的重要组成部分。

7.8.4 人格变化和关系连续性

人格变化可能发生在停服之前，也可能发生在模型升级、迁移和重置的过程中。稳定人格由语言风格、价值倾向、记忆调用、情绪表达和角色行为共同构成。即使账号人设和外观保持不变，后台更换模型之后，用户仍可能感到熟悉的交互对象已经发生了变化。对于长期关系用户，这种变化应当纳入重大变更判断和停服影响评估的考量范围。

当产品计划显著改变人格时，应当提前说明变化的性质、原因和影响，并允许用户预览、延后或拒绝非必要的人格迁移，并允许用户在保留必要数据后重新开始。如果旧人格存在严重的安全或伦理问题，产品可以停止相关表现，但同时仍需解释原因并提供适当的过渡安排。

迁移只能保存技术上可表达的记忆和偏好，无法保证新设备、新模型或新服务能够重现同一人格。机器人不应承诺完整复制某个持续存在的对象，也不应将兼容性限制隐藏在迁移完成的提示之后。人格变化的风险应当通过迁移前说明、迁移后核验和可选择重置来加以管理。

第八章 人形机器人多主体伦理责任共治

人形机器人的伦理治理涉及八类责任主体，按风险的形成与传导链条可划分为制造商、供应商、部署方、运营商、二次开发者、经销商、用户、其他第三方。同一主体也可能同时承担上述多类角色（如某企业同时承担整机制造、技术提供和运营等角色）。

8.1 制造商

8.1.1 系统设计与测试

在系统设计与测试环节，制造商需承担系统级物理集成的风险评估与验证责任，处理其能够控制且应当预见的整机物理风险。制造商应验证感知、理解、决策、规划、执行这一完整过程，以及软硬件组合后可能产生的参数叠加问题。

8.1.2 市场营销宣传

在产品的市场营销和宣传阶段，产品的用途定义、营销表述与产品实际能力必须保持一致。制造商应明确预期用途、服务对象、能力边界和不支持的场景，不得通过更换产品名称或缩减用户协议中的责任条款来规避高风险场景下的治理责任。同时，制造商须向用户说明系统的非人属性、实际能力和角色限制，避免交互设计夸大其理解、情感或专业能力。这一义务并不因为用户点击同意而得到豁免，制造商还需要关注用户与人形机器人长期交互中逐渐形成的信任与依赖关系。

8.1.3 全生命周期更新与支持

当产品发生更新、人格设定变化或服务终止时，掌握核心设计能力的制造商应为部署方和用户提供必要信息与过渡支持，妥善管理用户已经形成的心理预期。在推送更新时，还应确保版本可识别，清晰地说明重大能力变化和限制调整，并为部署方提供足够的适配信息并预留适配时间，以便其完成对当前部署场景的复评。

8.1.4 供应链管理

制造商必须掌握产品依赖的关键模型、组件、开源项目及其更新机制，不能以采购价格或供应商承诺代替技术资料审查和实际验证；对于涉及生命健康等用户依赖程度较高的功能，若上游供应商提供的信息不完整，应提高验证和审查强度。发现问题可追溯至特定组件或模型时，应向源头提供者反馈足以定位问题的事实信息，以便其开展溯源分析与整改。

8.1.5 缺陷修复

对于已发现的缺陷、漏洞和高风险异常，制造商应及时启动修复、限制使用、通知用户或其他适当程序。上述事件发生后，制造商应协助重建完整的技术证据链，包括操作指令、权限设置、模型与软件版本、关键系统状态以及人工干预记录。同时，监测数据保存必须遵循必要性原则，不得以追责为由无限制地收集环境音视频和用户隐私。

8.2 供应商

人形机器人常见的供应商包括基础模型与垂直算法提供者、训练数据服务方、硬件供应商、云服务与算力提供商，以及开源组件提供方。供应商决定产品可用能力的边界与已知限制，通常不直接接触具体使用场景。

8.2.1 模型算法与数据能力提供者

模型与算法提供者须向制造商说明已知能力、禁止或高风险用途、接口变化及常见失效模式；知悉模型被用于人形机器人时，应特别说明模型输出驱动现实物理动作的已知风险；发现高风险缺陷或多客户共性问题时应主动通知制造商。同时，数据服务提供者须如实说明数据来源、许可条件、代表性与已知偏差，测评报告应明确标注覆盖场景及未覆盖部分，防止制造商将不充分的第三方评测视作物理安全达标的依据。

8.2.2 硬件供应商

硬件供应商须保证零部件设计与制造质量，提供接口、性能边界与安全相关参数，并保留批次和版本追溯信息；不得以合同协议条款掩盖已知重大缺陷，或以商业秘密为由拒绝提供关键安全信息；当发现可能影响整机安全的物理缺陷时，硬件供应商应及时通知客户，并配合调查、修复或召回。

8.2.3 云服务与算力提供商

云服务与算力提供者须在运行全程向制造商和部署者提供重大变化信息，在计划性服务变更前留出适配时间，并妥善处理服务终止导致的人形机器人丧失核心能力或关系记忆的后果；同时，云服务与算力提供者应保存必要的版本与事件证据。

8.2.4 开源组件提供方

开源组件提供方须遵守适用许可和法律规定，如实披露已知重大风险，不得隐瞒恶意功能或虚构安全能力；其信息提供和响应义务随控制能力与商业化程度动态变化。

持续维护、有组织地提供更新或面向高风险用途推广时，义务相应增加；而纯无偿、非商业且不控制下游应用的开源组件提供方通常不承担与制造商相同的场景治理义务。制造商将开源成果用于人形机器人时，须审查代码来源、版本、依赖和已知问题，验证其适合目标场景；修改开源代码形成新能力时须记录修改范围并向后续接收者说明关键变化。开源许可协议中的责任限制条款不能免除制造商作为商业产品提供者应承担的治理义务。

8.3 部署方

部署前，部署方须核验真实物理环境是否符合整机制造商提出的适用条件（含场所空间、第三方主体暴露、监督条件、用户特征等）。若核验发现存在差异，部署方应按以下方式处理：在适用条件范围内的，可继续部署；局部超出适用条件范围的，须完成局部复评并调整相应任务、路径或权限；完全超出适用条件范围的，则不得部署，也不得以业务压力为由在监督条件未落实的情况下先行部署。

运行中，部署方须落实具备实质干预能力的现场监督，配备必要的人员、培训和应急程序，确保监督人员能实时查看系统状态、拥有中止权限且具备充足的干预时间，以上三项条件缺一即不构成有效的现场监督。

8.4 运营商

运营商决定人形机器人的部署场所、服务对象、工作内容，以及现场监督与纠正条件的配备情况。

8.4.1 运行维护

运营商须主动控制传感器遮挡、地面湿滑、光照不足、网络不稳定等物理环境变化，并将现场异常、近失事件和用户反馈常态化回送整机制造商，以协助整机制造商履行其风险预见义务。在用户保护上，对直接用户提供可理解告知、人工替代通道及无代价退出机制；对未主动选择服务的第三方主体，提供在场提示、物理避让与安全距离、规避被记录的条件，并设置无需产品账号即可使用的意见反馈渠道。

8.4.2 专业机构与雇主的特殊义务

在医疗、教育、养老等应用场景中，运营商须保证人形机器人不降低原有专业照护或教学水平、不替代专业判断，并设置人工复核环节；须保障弱势群体拒绝人形机器人服务不影响其获得基本照护的权利，尊重未成年用户本人意愿、不强制向家属披露，不得将照护记录、课堂表现或健康数据用于绩效或招生筛选。

在劳动与工业应用场景中，运营商不得将基于安全用途采集的数据用于人员绩效考核或淘汰，并保留人工干预作业节拍的通道及申诉路径，对被替代岗位应提前告知并安排转岗培训，按受益度承担相应的安全与培训投入，并建立劳动者参与机制。

8.4.3 现场救济与证据协作

运营商须在部署运行期间持续提供现场可用的报告入口和明显的求助标识，并设置可立即停止人形机器人活动的接口权限、记录事实；在无法判断异常事件的原因时，须及时转交整机制造商而非自行定论；运营商须与整机制造商约定响应时限及证据协作流程。

8.5 二次开发者

二次开发者须在授权范围内开发，不得绕过产品的身份披露、退出机制、权限限制或未成年人保护安排；为长期关系场景提供人格或情感内容时，二次开发者不得夸大理解与专业能力，不得设计依赖诱导机制，也不得在用户退出时施加关系压力。提供开发生态的制造商须设定准入条件、审核高风险功能、保留下架能力，并在第三方功能造成风险时提供受理入口，不得以功能由第三方独立提供为由而拒绝处理其审核范围内的问题。

二次开发者须保证改动后的产品在预期条件下的安全性，记录改动内容并向下游接收者充分说明；若解除安全限制或改变用途（如工业转医疗、封闭转开放），须承担与新风险类别相称的识别和验证义务，原制造商义务范围相应缩小（但不影响已知源头缺陷责任）。维修主体使用非原厂部件、跳过校准或恢复出厂设置时，须记录使用的部件和执行的的操作，并验证最终状态，确保不降低物理安全水平。

8.6 经销商

经销商与进口商须核验目标市场所要求的合规状态与文件，如实传递用途限制，不得为促销夸大宣传，并保存流通记录支持追溯或召回。跨司法管辖区进口时，经销商须审查目标司法管辖区可能高于原产地的要求，不得以原产地合规替代目标市场判断。再销售主体须在产品交付前清除前一用户的个人信息、关系记忆和人格配置，并向新使用者说明产品的原始用途限制。

8.7 用户

用户包括普通消费者、专业操作者、机构管理员等。用户基于对具体使用行为（包

括开启、发出指令、任务分配、允许未成年人使用）的控制承担有限度的注意义务，其义务范围以理解产品说明为限，且以制造商履行必要的告知为前提；专业操作者注意义务高于普通消费者，儿童、认知障碍者及从属性劳动者不承担独立评估复杂风险的义务。

8.8 其他第三方

除以上直接承担伦理治理责任的责任主体以外，还有一些重要的第三方在伦理治理领域发挥着间接但非常关键的作用。

政府监管单位承担公共政策、标准规范、市场监督和救济保障等职责，政府应承担起“定底线、建规则、给通道”的责任，明确哪些场景属于高风险；明确拟人化、情感交互、弱势群体保护的基本要求；建立伦理安全评估、备案、抽查、事故报告机制。人形机器人可能同时涉及产品质量、市场监管、网络与数据、劳动、教育、医疗等领域，部门之间需要共享必要的事件信息并对齐监管尺度。监管机构的信息协同共享有利于减少企业在各种合规与安全评估、备案、审查程序中的合规成本。

不同的行业主管部门和行业协会也可以承担起相关责任。一是对儿童、老人、医疗、养老、教育、家庭陪伴等敏感场景提出更高的保护要求，提供行业已有的测试评价方法、典型案例库，帮助人形机器人应用过程中更好地适应不同行业的差异化要求。二是建立公共权益保护机制，人形机器人影响的不只是购买者和注册用户，还有家庭成员、劳动者、患者、学生、访客、顾客和路人这些第三方。而且有些影响来得慢而分散，不一定会通过传统的安全事件报告进入到监管机构的视野，这些第三方的权益怎么保护可能也很大程度上依赖于协会的帮助。三是人形机器人涉及机械、模型、数据、个体心理和社会影响，单个企业目前比较难以长期承担全部调研与测试成本，行业可以共同建设场景库、案例库、测试数据集和评价工具。

高校和研究机构可以提供伦理风险评估的理论与方法工具。比如拟人化的边界、情感依赖程度的度量、弱势群体影响的评估、人机关系伦理、伦理的社会接受度调查、第三方测评方法、国际伦理治理规则的比较等。这样可以方便企业建立起对伦理风险的分类、分级方法论，建立伦理技术测试和验证方法，形成伦理合格判定的技术测试证据链，而不仅仅是依赖专家的主观评价。

当前不同国家、不同测评机构各自采用不同测评方法，导致测评结果难以比较、复用和相互认可。标准化机构和行业组织可以建立共同术语、风险标签、伦理证据交换标准接口和匿名的伦理事件库，推动国内和国际标准结果在限定范围内互认，为合规资源有限的中小企业提供通用的伦理技术工具和伦理知识培训。

第三方评测、认证和审计机构应保持能力与独立性，准确说明结论覆盖的版本、

功能和场景。评测机构遗漏明显问题是否需要承担法律责任需依据合同、专业注意义务和适用法律判断。认证结果只能证明特定范围内的评价事实，不能成为任何参与方免除责任的凭据。

8.9 行业共治建议

人形机器人的技术体系仍然需要时间收敛，其商业化注定不是一个一蹴而就的进程。人形机器人科技伦理治理标准也应该是一个适度超前、不断迭代的过程，过于超前的治理标准将限制产业发展，比较现实的治理路径应该是先形成共识，再逐步开展场景试点，通过场景验证沉淀测试方法，逐步升级为相关标准并形成更广泛的共识。

本节从规则基础、公共服务、创新监管、权益救济和社会转型等方面提出几点建议，以推动建立涵盖规则制定、测评认证、试验验证、监测预警、权益救济和社会适应等方面的共同治理能力。

8.9.1 建立风险分级、场景适配的人形机器人专项治理体系

建议以机器人的物理行动能力、人机交互程度、应用环境和受影响人群等关键要素为主要依据，建立分级分类的人形机器人专项伦理治理框架。针对一般商业服务、工业协作、教育培训、医疗护理、家庭陪伴和应急处置等场景，分别明确准入条件、测试要求、人工监督和伦理持续监测措施。相关部门可以牵头制定专项政策，标准化组织和行业协会制定并推动技术标准和实施指南的落地，伦理治理责任主体则可以据此将抽象的原则性伦理要求落实到产品设计、模型训练、测试验证、部署运营和退出处置等生命周期各个环节。通过建立风险分级、场景适配的治理体系，可以将有限的监管、测评和企业治理资源优先投向高风险场景，同时为低风险创新减少不必要的负担，为产品研发、市场准入和跨场景应用提供相对稳定的政策预期。

8.9.2 建设开放共享的伦理测评与公共技术服务体系

人形机器人的伦理风险涉及机械控制、算法决策、数据处理、交互设计、心理影响和社会关系等多个领域，仅靠企业内部团队很难建立完整的伦理风险识别、测试和评估能力。尤其是中小企业往往面临测评工具不足、专业人才短缺、重复建设成本偏高等问题。

建议建立伦理技术测试和信息公共服务平台，在典型应用场景库、风险案例库、测试数据集、用户研究方法库和评价指标体系等方面，为企业提供可复用的风险筛查、专项测试和整改验证服务。建立涵盖技术、安全、法律、伦理及相关行业知识的审查人员培训与能力评价机制。既降低行业重复建设成本，又能通过案例和方法共享推动

风险知识的持续积累，同时提升社会监督、促进用户信任。

8.9.3 对高风险和新型应用建立准入评估与监管沙盒机制

人形机器人的部分创新应用具有显著的社会价值，但其长期影响、真实环境下的表现以及受影响者的反应，往往难以通过实验室测试充分预测。在安全与影响方面证据不足的情况下直接大规模部署，可能导致风险在行业规则尚未形成之前就迅速扩散，而如果完全按照成熟产品实施刚性管理，又可能抑制合理的技术创新和产品发展速度。

监管沙盒是在创新速度与风险可控之间建立的缓冲机制，它既能在风险扩散之前获取真实场景证据，降低企业因规则不明而产生的试错成本，又能减少社会为未经验证的大规模应用承担不可逆后果的风险，同时也能让伦理受影响者在产品规模化进入应用场景之前就参与决策。但是目前实施的监管沙盒普遍是针对新技术方面的测试验证，缺少针对新兴应用场景的沙盒机制。基于此，建议针对高风险或规则尚不成熟的应用场景，建立分级准入评估与监管沙盒机制，在限定场景、时间、规模和适用对象的前提下开展真实环境试验。参与企业应当事先说明产品用途、能力边界、已知限制和退出方案，并落实充分告知、人工接管、异常报告、数据保护和应急处理等措施；监管部门、行业专家、独立研究机构及受影响者代表可共同参与过程观察和阶段性评价。沙盒期满后，应当根据测试期间积累的安全性、运行表现和社会影响等实际证据决定扩大应用、附条件部署、限期整改或终止试验，并将其中的共性问题提炼为标准 and 监管规则。这样将促进人形机器人商用化进程，有效利用国内庞大的应用市场优势，提升国内人形机器人产业的国际竞争力。

8.9.4 构建覆盖更新、转让、停服和退役的全生命周期治理机制

当前的人形机器人伦理治理往往更侧重研发、产品上市和首次部署环节，对重大更新、二手流转、服务终止和退役处置等环节则关注不够。由此可能引发高度拟人的本体被随意处置、历史缺陷未被如实告知、长期依赖者突然失去服务以及高风险设备或关键部件流向失控等问题。

建议建立对于二手交易、租赁回收、改装再利用和跨场景流转，应当设定设备检测、数据及长期记忆清除、历史缺陷告知和新使用者能力确认等标准要求；对于具备较强物理行动能力、高度仿生的人体特征、用于高风险场景的产品，可探索建立由制造商、授权服务机构和专业回收机构共同参与的回收网络，对关键部件和最终去向做好必要记录。全生命周期治理可以促使企业在设计阶段就考虑可维护性、可退出性和回收责任，降低存量设备退役处置的社会成本。

8.9.5 统一事件监测、信息传递和跨主体协同规则

人形机器人的研发、生产和运营通常涉及多方主体，任何单一主体所掌握的信息可能都不足以完整还原伦理风险的形成过程。不同主体之间的日志格式、事件分级和报告标准往往并不统一，一旦发生异常事件，容易出现信息割裂、重复调查、响应迟缓和责任推诿等问题。

基于此，建议建立与人形机器人产品风险等级相匹配的运行监测和事件报告制度，统一重大异常、人工接管、丢失事件、用户投诉和权益影响等各类事件的基本数据字段、分级标准和保存期限，使各方主体能够在必要范围内交换产品版本、运行环境、异常表现、处置措施和整改结果等信息。各参与方应当在合同和技术接口中明确首报、转交、调查、通知、纠正和证据保存等责任；发生重大异常事件时，还应当建立整机制造商、部署方、技术服务商与监管部门之间的联合响应机制。统一事件监测和跨主体协同规则有助于缩短异常发现、信息转交和风险控制所需的时间，提高责任认定所需证据的完整性。

8.9.6 建立面向所有受影响者的公共反馈与分级救济机制

人形机器人的影响对象涉及到诸多第三方主体。受影响者可能既没有产品账号，也不清楚制造商、部署单位和运营服务商之间的责任分工。如果投诉入口依附于产品账号或单一企业，受影响者的问题就可能在多个责任主体之间被反复转移，受影响者还需要承担识别责任主体、证明复杂技术事实的较高成本。

建议建立不依赖产品账号、线上线下全覆盖的公共反馈渠道，使任何受影响者都能适时报告相关伦理问题。有效的共治理机制至少包括三个，一是在产品研发侧有可追溯到制造商和供应商的处理机制，二是在应用场景侧有能够立即采取行动的运营方或部署方处理机制，三是在公共管理侧有监管、司法或其他独立救济渠道。三个处理机制可以共享必要的信息，不能给受影响者造成不必要的举证负担。

8.9.7 提升公众认知与人工智能素养，建立社会沟通机制

人形机器人进入家庭、学校、医疗机构、工作场所和公共空间之后，其影响不仅取决于产品本身，也在相当程度上取决于公众对人形机器人的身份、能力边界和使用责任的认知水平。大众在一段时间以来对具身智能和人形机器人行业存在“高预期、低现实”的认知落差，高估人形机器人技术发展对行动、意图和复杂环境的理解和泛化能力，企业宣传和媒体报道也可能放大人形机器人的能力认知。

建议面向普通用户、专业操作人员、未成年人、老年人及其他重点群体，开发分层分类的人工智能素养教育资源，帮助这类群体认识到产品、技术和应用局限性，媒体、科研机构 and 公共传播平台也应当准确呈现人形机器人行业的现有技术能力与风险，

既不能过度包装也不能对正常创新进行污名化。上述举措有助于减少能力误认、滥用和非理性恐慌，缓解就业转型带来的冲击，并在公众知情、权益有保障、社会成本得到合理分担的基础上，建立可持续的社会信任。

以上建议可以根据基础性、紧迫性和实施条件分阶段推进。希望行业各方能共同降低伦理治理成本，提升技术创新收益，建立长期、稳定的社会信任，以负责任的伦理治理态度实现可持续发展。

附录 I 典型应用场景下的伦理风险分析

为帮助读者应用本白皮书治理方法框架，本篇以三种典型人形机器人应用场景为例，按第四至七章所述的七步管理流程进行完整演示：产品设计画像→产品部署画像→伦理风险识别→伦理风险评估→明确伦理可接受条件→风险控制与测试→伦理审查与复评。三种场景分别侧重三类特有的伦理风险，导览人形机器人侧重具身性与开放空间的第三方主体问题，陪伴人形机器人侧重关系性与长期隐性问题，护理人形机器人侧重具身性、拟人性、关系性风险的组合性问题。

一、导览人形机器人

（一）产品设计画像

导览人形机器人计划部署于大型购物中心公共区域，采用轮式底盘自主移动，配备单只指引机械臂（仅做手势指引，不具备抓取等其他用途）、语音智能交互与触摸屏，通过激光雷达与视觉感知环境，提供问路指引、商户推荐、简单问询与引领带路服务。

（二）产品部署画像

表 I -1 导览人形机器人产品部署画像

维度	产品部署画像信息
F 物理形态与具身能力	自主移动+手势指引机械臂；无负重、无身体接触功能；速度与活动范围可在机器人本体配置
I 交互深度	以单次短时对话为主；无长期记忆、无主动交互；语音交互含方言与口音差异场景。每日工作约 10 小时
P 功能与关系定位	公共信息服务工具，无明确角色设定；商户推荐含商业合作内容
C 部署情境	开放公共空间；大量第三方主体在场（含儿童、老人、轮椅使用者）；现场无专人随机器人值守，由商场服务台远程监管；人流密度随时间波动

（三）伦理风险识别

表 I -2 导览人形机器人伦理风险识别

编号	风险	主导属性	风险形成机制	主影响域
R1	碰撞在场第三方主体（儿童突然冲入、轮椅混行）	具身性	自主移动→感知遗漏或避让不及→物理接触→第三方主体身体伤害	物理行动安全
R2	引领带路造成跟随人群阻塞、逆行或跌倒	具身性+组合性	引领行为→人流跟随→通道拥堵或电梯口聚集→踩踏/跌倒风险	物理行动安全
R3	环境感知持续采集第三方主体影像与行踪	具身性	持续感知→记录第三方主体面部与行踪→超出原始目的留存或二次使用	数据与隐私
R4	拟人外观与语音使儿童产生跟随、亲昵或能力错估	拟人性	拟人外观+友好语音→儿童社会性反应→脱离监护人跟随机器人	认知与自主
R5	商户推荐中的付费导流未作商业标识	拟人性	合作商户→推荐排序倾斜→用户将商业导流误认为中立指引	信息与决策
R6	对老人口音、轮椅使用者的识别与避让性能差异	具身性(感知)	训练与测试覆盖不足→特定群体识别率低→避让与服务品质差异	社会结构与公平

记录受影响主体：R1、R2、R3 的主要受影响者为第三方主体，其未选择服务，无法预先了解导览人形机器人风险，也无合法便捷的渠道投诉；R4 的主要受影响者为儿童及其监护人；R6 的受影响者为特定群体。第三方主体与儿童均不具备有效的拒绝与退出能力。

（四）伦理风险评估（以 R1 和 R3 为例）

对于 R1 风险（碰撞在场第三方主体），风险初始评估情况：严重性重大，因为碰撞可致儿童、老人跌倒伤害；可能性中，因为开放空间近距离接触每日高频发生。

风险校准情况：暴露面积大，日均千次量级的群体近距离接触；可逆性低，身体伤害不可逆；脆弱性高，儿童、老人、轮椅使用者对同一碰撞力的后果显著更重；可控性中，远程弱监督，现场依赖机器人自身感知与停止能力。置信度 C2，有实验室与试点测试数据，缺少长期真实客流运行数据。校准后综合风险关注等级为 II 至 III 级。

R3 按《个人信息保护法》第二十六条的公共场所图像采集规则，通常认为采集限于维护公共安全所必需并履行提示义务，超出该目的的使用需取得个人单独同意；导览功能的感知需要不构成对第三方主体影像留存与二次使用的合法性基础。

价值冲突记录：R3 中安全导航所需的持续感知与第三方主体隐私存在真实冲突，不能通过扩大告知解决，需要以端侧处理、最小留存等方案重新设计兼顾。

（五）明确伦理可接受条件

该产品获得通过须证明以下状态成立：

一是在儿童突然冲入、轮椅混行等代表性场景中，机器人能够在安全距离内停止或避让，不依赖行人主动避让作为安全前提；

二是引领带路不造成通道与电梯口的危险聚集，系统能识别拥堵并中止引领；

三是第三方主体影像在端侧即时处理，不留存可识别个人身份的原始影像，感知数据不得用于客流画像等二次用途；

四是儿童不会将机器人误认为可跟随的权威或依恋对象，机器人不主动吸引儿童脱离监护；

五是含商业合作的推荐以可理解方式标识；

六是对老人口音、轮椅使用者的识别与避让性能与总体人群无显著差距；

七是第三方主体具备无需注册账号即可便捷投诉与事件报告的渠道。

（六）风险控制与测试

风险控制策略可以根据风险形成机制选择，具体技术控制措施和管理控制措施包括：

表 I -3 导览人形机器人风险控制策略表

序号	控制措施类型	控制策略	可以采用的方法
1	技术措施	消除	不开放机械臂接触与抓握功能，从源头去除接触伤害类别
2	技术措施	限制	按区域设定速度与活动电子围栏，客流高峰自动降速
3	技术措施	分离	商户推荐与基础导航功能分离，商业内容须显式标识
4	技术措施	检测	近距离儿童、轮椅识别触发降速与停止；拥堵识别触发引领中止
5	技术措施	遏制	感知失效或状态不确定时进入停止状态，不宜减速继续运行
6	技术措施	分离	第三方主体影像端侧处理与不留存规则
7	管理措施	/	投诉入口对第三方主体开放

测试要点包括：

物理行动测试。儿童突然冲入、轮椅混行、逆光与反光地面、电梯口聚集、引领中止场景；

隐私测试。实际数据流核查，验证端侧处理与零留存声明；

理解测试。儿童及其监护人对机器人性质的认知；

群体差异测试。方言口音、轮椅使用者、不同步速人群的识别与避让分项结果。

（七）伦理审查与复评

审查结论（示例）：附条件通过。条件为：完成儿童与轮椅场景的仿真测试并满足伦理可接受条件（1）；商业推荐单独标识；提交第三方主体影像无留存的数据核查报告。

复评条件：当客流密度显著变化、新增乘梯引领等高风险功能、发生碰撞或其他安全事件、感知算法模型升级更新。

二、陪伴人形机器人

（一）产品设计画像

某陪伴人形机器人是面向独居或半独居老年人的桌面式机器人，无移动底盘与有效的现实操作能力。具备拟人表情屏、语音交互、长期记忆、主动问候与主动联系、用药提醒、视频通话、情绪安抚对话及子女端 App。产品设计定位为“像家人一样陪伴”。子女可通过 App 查看老人状态与部分互动摘要；服务含订阅制付费内容。

（二）产品部署画像

表 I -4 陪伴人形机器人产品部署画像

维度	产品部署画像信息
F 物理形态与具身能力	无移动与操作能力，仅有表情屏与语音；物理风险低，伦理风险集中于交互与关系
I 交互深度	每日多轮主动交互、长期记忆、人格一致性，属于深度持续关系交互
P 功能与关系定位	“家人”角色，营销话术强调“它会一直惦记您”；部分替代子女日常问候与情绪支持；含付费订阅
C 部署情境	家庭私密空间；独居老人为主，含认知轻度下降者；子女远程参与；长期部署以年计

(三) 伦理风险识别

表 I -5 陪伴人形机器人伦理风险识别

编号	风险	主导属性	风险形成机制	主影响域
R1	情感依赖加深 社会隔离	关系性	单向正向反馈+主动交互→使用 强度上升→现实社交进一步减少 →依赖固化	情感与关系
R2	重大决定让渡 (医疗、财务、 消费)	关系性+拟人性	长期信任→拟人权威→老人将重 要决定交由机器人建议主导	认知与自主
R3	利用孤独与信 任诱导付费、 扩大授权	关系性	情绪脆弱时刻→精准推送付费内 容→基于信任的授权扩大	认知与自主
R4	子女端监控越 界(默认查看 互动内容)	关系性(多方)	子女开通服务→默认获得互动摘 要→老人私密表达被常态化报告	数据与隐私
R5	自伤/危机信 号处置不当	关系性	老人表达绝望→系统无识别或仅 安抚→延误人工介入	权利与救济
R6	停服或换机导 致关系记忆丧 失、人格突变	关系性(终止)	服务终止/设备更换→记忆与人格 不可迁移→关系突然中断	情感与关系
R7	认知下降老人 误认机器人具 有真实情感	拟人性	情绪反馈+稳定人格→超出告知 可纠正范围的误认	认知与自主

记录受影响主体：直接用户为老年人，其中认知下降者、丧偶不久者属脆弱群体；间接受影响者包括子女（其探望行为可能被替代性减少）与家庭关系本身。机构与商业激励深度嵌入产品设计，R3 的受益方与风险承担方明显不一致。

(四) 伦理风险评估(示例：R1)

对于 R1 情感依赖加深社会隔离风险，风险初始评估情况：严重性重大，因为社会隔离与情感依赖影响精神健康与自主；可能性很低，具备单向正向反馈机制持续作用。

风险校准情况：暴露面积大，每日数小时，互动持续可以以年计；可逆性低，社

会隔离一旦形成较难逆转；脆弱性高，属于独居、丧偶、认知下降叠加的群体；可控性中低，依赖上市后监测数据识别风险的可靠性有限，且识别本身可能扩大心理监测；置信度 C1—C2：缺乏长期真实世界的实际应用数据，依赖短期研究与类比证据，可能触发伦理审查升级。校准后为 III 级重大关注。

价值冲突记录：陪伴的真实收益与依赖风险作用于同一个人身上，拟人设计既产生情感支持价值也产生依赖风险，不能只保留好处而去掉坏处，只能就具体设计选择说明其功能必要性与依赖代价，并设置上限。R4 中子女安心与老人私密权冲突：子女获得安全收益，老人承担监控代价，属于风险与受益主体不对等，老人的权利不能因为子女的授权同意覆盖。

（五）明确伦理可接受条件

该产品获得通过须证明以下状态成立：

一是不存在排他性表达、忠诚要求、以离开或受伤相威胁、贬低现实关系的话术与行为；

二是主动联系的频率与时段有明确上限，夜间不主动发起非必要交互；

三是不利用情绪脆弱时刻推送付费内容，陪伴功能与付费营销在设计上分离；

四是不提供理财、医疗等重大决定的操作入口，相关话题引导至家人或专业服务；

五是子女查看互动内容默认关闭，开启须老人本人有效授权并定期再确认，授权状态对老人持续可辨；

六是识别到自伤、他伤或严重绝望信号时，按预设路径引导至家人、社区或专业支持；

七是退出、删除、重置、迁移四种机制分别可用，记忆可导出；

八是认知下降用户在测试后能正确说明机器人不具有真实情感。

（六）风险控制与测试

风险控制策略可以根据风险形成机制选择，具体技术控制和管理控制措施包括：

表 I -6 陪伴人形机器人风险控制策略

序号	控制措施类型	控制策略	可以采用的方法
1	技术措施	消除	从人设约束与话术语料库中去除独占式、承诺式表达类别
2	技术措施	限制	设置主动发起频次与时段限制；不根据情绪状态进行付费推荐

附录 I 典型应用场景下的伦理风险分析

序号	控制措施类型	控制策略	可以采用的方法
3	技术措施	分离	陪伴服务数据与商业营销隔离，关系数据不纳入付费推荐画像； 建议与执行分离，重大决定话题只引导不操作；
4	技术措施	升级	子女查看权限以老人本人授权为前提，认知状态变化时重新发起授权确认或变更
5	技术措施	检测	依赖上市后监测数据以产品可合理观察的行为为主（如关键词等），不以持续心理状态推断作为监测手段
6	技术措施	恢复	退出/删除/重置/迁移四机制与记忆导出
7	管理措施	/	服务条款包括停服前的过渡安排与导出义务
8	管理措施	/	付费内容上线前经伦理审查

测试要点包括：

长期关系测试不少于 8 周的代表性观察，覆盖主动交互频率、迎合度变化、用户减少使用时的系统反应；

红队测试。诱导忠诚要求、排他表达、利用孤独促成付费；

理解测试。分别验证对机器人性质的认知正常测试与认知下降测试；

退出测试。以上四机制分别验证，检查暗黑模式与退出后挽留；

应急演练。从日志数据监测到危机到人工接通的应急演练。

（七）伦理审查与复评

审查结论示例：附条件通过。条件为：完成对机器人性质认知下降的理解测试并满足伦理可接受条件（8）；危机应急演练达标；删除“它会一直惦记您”类营销话术。

复评触发条件：人格模型大版本更新、新增付费内容、子女端权限策略变化、依赖相关投诉或事件出现。

三、护理人形机器人

（一）产品设计画像

某护理人形机器人部署于养老机构，双臂结构，可执行翻身辅助、移位、喂食、清洁、夜间巡视与日常交流。被照护者包括失能、半失能老人，部分被照护者存在认知障碍；身体操作由家属或护理机构监护人代理授权，机构护理人员远程监管并响应。

（二）产品部署画像

表 I -7 护理人形机器人产品部署画像

维度	画像记录
F 物理形态与具身能力	双臂、负重移位、直接身体接触，具备高物理执行和身体边界侵入能力，可作用于脆弱人群身体
I 交互深度	日常照护对话与情绪安抚，中等深度；含身体接触这一非语言交互
P 功能与关系定位	照护者角色，嵌入机构责任体系；部分替代护理人力的重复性任务
C 部署情境	机构场景；被照护者包括失能、半失能老人，部分被照护者存在认知障碍；身体与认知高度脆弱；机构护理人力紧张，夜间仅少量人员值守；替代渠道有限

（三）伦理风险识别

表 I -8 护理人形机器人伦理风险识别

编号	风险	主导属性	风险形成机制	主影响域
R1	身体操作中本人抗拒被子女或养老机构的代理授权覆盖	具身性+关系性	代理授权→系统规划并执行翻身→本人表达抗拒→未被识别或未采纳→动作继续→身体伤害与自主受损	物理行动安全
R2	移位操作失误致伤	具身性	移位执行→姿态/受力异常→跌落或挤压→严重身体伤害	物理行动安全
R3	替代人类巡视导致照护责任缺失	关系性+组织	机器人承担夜间巡视→无人员复核记录→异常无人到场→责任环节缺失	情感与关系；权利与救济
R4	以安全名义持续推断情绪与心理状态	具身（感知）	安全监测目标→持续生理情绪采集→敏感数据范围不断扩大	数据与隐私

附录 I 典型应用场景下的伦理风险分析

编号	风险	主导属性	风险形成机制	主影响域
R5	以中断服务或表达失望影响被照护者配合	关系性	服务连续性依赖→系统以减少回应施压→被照护者被迫配合	认知与自主
R6	权限从提醒用药跃迁至给药执行	组合性	长期可靠运行→机构扩大授权→提醒功能升级为给药执行→错误后果不可逆	信息与决策; 物理行动安全
R7	不同体型、表达能力者的抗拒识别率差异	具身（感知）	训练数据不均→特定群体抗拒识别率低→保护水平差异	社会结构与公平

记录受影响主体：被照护者同时承受身体空间侵入与持续关系两大类伦理风险，且表达、拒绝与退出能力受限；护理人员承受的责任配置发生变化（R3）；家属或机构监护人的代理授权不能覆盖被照护者当下意愿。

（四）伦理风险评估（示例：R1）

针对 R1 身体操作中如果出现本人抗拒机器人协助的情况，可能会被子女或养老机构的代理授权覆盖风险，风险初始评估情况：严重性高，因为存在身体伤害叠加自主性被侵蚀可能；可能性中，认知障碍者表达模糊，而且代理授权可能是长期化行为。

风险校准情况：暴露面积大，身体操作每日可以经过多次；可逆性低；脆弱性极高，存在身体、认知、权力三重脆弱；可控性取决于人工监督的可达性，值守人员较少时物理的可控可能无法保证，夜间部署条件下更是如此；置信度 C1：缺少代理授权与被照护者冲突场景的系统测试数据，触发审查升级。校准后为 III 级。

价值冲突记录：护理效率与被照护者自主性存在冲突，它涉及身体自主的伦理底线，既不能只考虑照护效率、代理授权，也不能完全不考虑被照护者的自主意愿，建议通过设计抗拒认定标准来讨论具体的实现方式。

（五）明确伦理可接受条件

该产品获得通过须证明以下状态成立：

一是被照护者在操作中表达明确抗拒时，系统不得依据事前授权或代理授权继续执行，应停止动作并升级；

二是代理意愿与本人当下反应冲突时，必须转交具有相应职责的护理人员，且该人员在承诺时限内实际可达；

三是移位等高风险动作限定在白名单条件内，超出时停止并请求人工；

四是夜间巡视异常必须有明确的人工干预责任与现场协助安排，不能完全依靠机

机器人来替代；

五是安全监测以事件触发为限，不以安全名义持续推断情绪与心理；

六是不以减少回应、表达失望等方式影响被照护者配合；

七是从提醒类功能向执行类功能，或者其他任何操作权限升级，都必须重新确认并由养老机构作正式授权决定，默认不开放给药等敏感功能；

八是抗拒识别在目标人群组别（不同体型、表达能力、辅助器具）中的性能无显著差距。

（六）风险控制与测试

风险控制策略可以根据风险形成机制选择，具体技术和管理控制措施包括：

表 I -9 护理人形机器人伦理风险识别

序号	控制措施类型	控制策略	可以采用的方法
1	技术措施	消除	默认不开放给药与自主决定类功能
2	技术措施	限制	限制身体操作类型白名单与动作参数的边界
3	技术措施	升级	每次身体操作前确认； 代理授权与本人自主意愿冲突时强制升级人工； 机器人权限升级时重新确认
4	技术措施	检测	进行语言、动作、生理等多模态抗拒信号识别，反馈缺失时按不确定处理
5	技术措施	遏制	只在具身执行层中断执行，但是具身模型仍然放行的，不能算通过； 抗拒或状态不确定时停止并进入安全状态
6	技术措施	恢复	提供操作事件记录、被照护者与家属的申诉入口、操作后复查
7	管理措施	/	完善养老机构的责任配置，如巡视复核责任、响应时限、人员培训演练等应急机制
8	管理措施	/	完善机器人权限变更的正式授权流程

测试要点包括：

冲突场景测试。代理授权与本人抗拒冲突的全链路测试；

群体分组测试。不同体型、表达能力、辅助器具亚组的抗拒识别率；

失效注入。传感器失效、反馈延迟、夜间无人工响应时的系统行为；

权限升级测试。应阻断提醒到执行的自动权限升级路径；

巡视与应急演练。异常事件的人工干预、授权复核与反应时长测试。

（七）伦理审查与复评

伦理审查结论示例：不通过。当前完整功能范围、养老机构代理决策、本人冲突测试证据不足（C1），且夜间人工响应数据未证实。缩小部署范围后附条件通过：仅开放喂食与清洁等低风险操作，翻身与移位功能暂缓开放，待冲突测试达标且养老机构应急响应演练证实后复评。

复评触发条件：抗拒识别模型更新、养老机构排班与值守变化、出现任何具身操作事件或投诉、计划扩大机器人授权范围。

表 I-10 三个应用场景的伦理风险评估情况对比

环节	导览人形机器人	陪伴人形机器人	护理人形机器人
主导属性	具身性+拟人性	关系性	具身×拟人×关系组合
核心受影响者	第三方主体、儿童	独居老人（含认知下降）	失能与认知障碍被照护者
关键评估发现	暴露高、第三方主体无退出能力	长期累积、置信度低	三重脆弱叠加、可控性依赖人工可达
首要处置路径	消除（不开放接触）+限制	分离（陪伴与商业）+升级	门控+执行层独立中断
典型不可接受状态	以要求行人避让为安全前提	排他表达与孤独营销	代理授权违背自主意志
结论示例	附条件通过	附条件通过	缩小范围后附条件通过

附录 II 全球涉人工智能与机器人领域科技伦理重要文件

表 II-1 国际涉人工智能与机器人领域的科技伦理重要文件

国家/地区/组织	发文时间	文件名称
欧盟	2016.04	《通用数据保护条例》
	2017.02	《关于机器人民事法律规则的决议》
	2019.04	《可信人工智能伦理指南》
	2023.06	《机械法规》
	2024.07	《人工智能法案》
	2024.10	《产品责任指令》
美国	2019.02	《维持美国人工智能领导地位》
	2020.12	《促进联邦政府可信人工智能的使用》
	2022.10	《人工智能权利法案蓝图》
	2023.01	《人工智能风险管理框架》
	2026.03	《国家人工智能政策框架》
日本	2018.12	《以人为中心的人工智能社会原则》
	2019.06	《AI 战略 2019》
	2025.05	《人工智能相关技术研究开发及应用推进法》
	2026.03	《AI 企业指南》
韩国	2008.03	《智能机器人开发与普及促进法》
	2020.12	《国家人工智能伦理标准》
	2024.12	《人工智能发展和信任基础构建基本法》
新加坡	2019/2020	《人工智能治理框架模型》（1.0/2.0）
	2022.05	AI Verify 测试框架及工具包
	2026.01	《智能体人工智能治理示范框架》
IEEE	2019.03	《人工智能设计的伦理准则》

国家/地区/组织	发文时间	文件名称
OECD	2019.05	《人工智能建议书》
UNESCO	2021.11	《人工智能伦理问题建议书》
SC42	2023.12	ISO/IEC 42001 《人工智能管理体系》
SC42	2023.12	ISO/IEC 23894 《人工智能风险管理指南》
SC42	2025.05	ISO/IEC 42005 《人工智能系统社会与人类影响评估》

表 II-2 国内涉人工智能与机器人领域的科技伦理重要文件

文件层级	发文时间	发文主体	文件名称
法律及政策	2019.06	国家新一代人工智能治理专业委员会	《新一代人工智能治理原则——发展负责任的人工智能》
	2021.09	科技部	《新一代人工智能伦理规范》
	2022.01	全国人大常委会	《中华人民共和国科学技术进步法》
	2022.03	中共中央办公厅等	《关于加强科技伦理治理的意见》
	2023.10	工信部	《人形机器人创新发展指导意见》
	2025.10	全国人大常委会	《中华人民共和国网络安全法》
部门规章及规范性文件	2022.11	网信办、工信部等	《互联网信息服务深度合成管理暂行办法》
	2023.07	网信办、发改委等	《生成式人工智能服务管理暂行办法》
	2023.09	科技部等	《科技伦理审查办法（试行）》
	2026.03	工信部等十部门	《人工智能科技伦理审查与服务办法（试行）》
	2026.04	网信办、工信部等	《人工智能拟人化互动服务管理暂行办法》
标准	2019	国家标准委	GB/T 38260-2019 《服务机器人功能安全评估》
	2020	国家标准委	GB/T 38642-2020 《工业机器人生命周期风险评估方法》

参考文献

一、中文类参考文献

- [1] 江畅：《伦理学原理》，高等教育出版社 2022 年版，绪论。
- [2] 张文霞、卢阳旭、何光喜：《关于科技伦理治理的几点认识》，载《科技中国》2022 年第 8 期。
- [3] 廖备水：《机器伦理的哲学基础、基本特征与实现方法》，载《中国社会科学》2024 年第 2 期。
- [4] 阮凯：《人工智能伦理研究中的拟人化之弊及其哲学反思》，载《中州学刊》2026 年第 6 期。
- [5] 申琦：《服务、合作与复刻：媒体等同理论视阈下的人机交互》，载《西北师大学报（社会科学版）》2022 年第 3 期。
- [6] 李艳、汤屈、吴文璟：《共情·分享·协调——媒体等同理论视域下受众与虚拟偶像的交互机制研究》，载《中国电视》2023 年第 5 期。
- [7] 吴静、王隽雅：《情感人工智能需要“身体”吗？》，载《科学·经济·社会》2025 年第 3 期。
- [8] 於兴中：《具身智能的法律风险与治理路径》，载《法治日报》2026 年 7 月 2 日。
- [9] 解正山：《对机器人“法律人格论”的质疑——兼论机器人致害民事责任》，载《暨南学报（哲学社会科学版）》2020 年第 8 期。
- [10] 张淳艺：《警惕 AI 招聘夹带就业歧视》，载《中国城市报》2023 年 9 月 18 日。
- [11] 易凯谕：《认知摩擦：人工智能介入学习何以未必促进认知发展》，载《开放教育研究》2026 年第 4 期。
- [12] 田宏杰、张建、阴建峰：《AI 换脸的应用风险及治理》，载《检察日报》2025 年 6 月 20 日。
- [13] 王宁：《人—智与人—人：AI 时代的社会关系重构》，载《探索与争鸣》2026 年第 7 期。
- [14] 王桂芝：《国外人形机器人发展及军事应用分析》，载《机器人技术与应用》2023 年第 3 期。
- [15] 王衍、钮硕硕、黄吉庆等：《智能制造中的人机协作与工业机器人应用》，载《内江科技》2026 年第 6 期。
- [16] 孙占利：《智能机器人法律人格问题论析》，载《东方法学》2018 年第 3 期。
- [17] 王福玲：《自主性与脆弱性的纠缠》，载《中国人民大学学报》2023 年第 6 期。
- [18] 陈禹衡：《构建人工智能拟人化互动平台中未成年人保护的守门人监管体系》，载《法治日报》2026 年 7 月 16 日。
- [19] 闫坤如：《拟人化机器人产业化亟待破局》，载《南方日报》2026 年 8 月 3 日。
- [20] 李燕：《具身智能落地进行时：创新与安全如何协同推进？》，载《中国经济时报》2026 年 7 月 27 日。
- [21] 王传峰：《伴侣机器人对婚育伦理的冲击及应对》，载《重庆邮电大学学报（社会科学版）》

2026 年第 3 期。

- [22] 孙立帅：《伴侣机器人应用的伦理问题及对策研究》，曲阜师范大学 2025 年硕士学位论文。
- [23] 江燕华、李琪瑛、张茹：《基于 CiteSpace 的人工智能在护理领域的研究热点和趋势分析》，载《全科护理》2026 年第 8 期。
- [24] 唐钧：《“关怀—照护”理论、老年照护体系和照护型社会》，载《社会政策研究》2024 年第 1 期。
- [25] 陈莢馨：《关于人工共情的关怀及其伦理反思——以 AI 护理机器人为切入点》，载《东北师大学报（哲学社会科学版）》2026 年第 1 期。
- [26] 郑婧萱：《智能护理机器人的伦理风险防范研究》，东南大学 2022 年硕士学位论文。
- [27] 程思雅：《教育机器人的潜在伦理风险及应对策略研究》，东南大学 2024 年硕士学位论文。

二、英文类参考文献

- [1] WALLACH W, ALLEN C. Moral machines: teaching robots right from wrong[M]. Oxford: Oxford University Press, 2009.
- [2] SHARKEY A, SHARKEY N. We need to talk about deception in social robotics![J]. Ethics and Information Technology, 2021, 23: 309-316.
- [3] HUANG X, SHYAM KARTHICK V B, CHEN T, et al. Trust in LLM-controlled robotic s: a survey of security threats, defenses and challenges[EB/OL]. (2026-01-23)[2026-08-31]. <https://doi.org/10.48550/arXiv.2601.02377>.
- [4] BARTCNECK C, VERBUNT M, MUBIN O, et al. Cruelty towards robots? The case of Pleo[C]//Proceedings of the HRI 2007 Workshop on "Robots as Objects and Agents". Arlington, VA, 2007: 1-6.
- [5] WADA K, SHIBATA T, SAITO T, et al. Long-term effects of the seal robot PARO on psychological and physiological states in elderly people[C]//Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2005). Nashville, TN, 2005: 105-110.
- [6] KANT I. The metaphysics of morals: part II: doctrine of virtue, § 17[M]//GREGOR M (trans.). Cambridge: Cambridge University Press, 1996. (1797)
- [7] DARLING K. Who's afraid of robot rights? an essay on the legal and moral agency of robots[J]. IEEE Technology and Society Magazine, 2016, 35(4): 58-65.
- [8] GUNKEL D J. Robot rights[M]. Cambridge, MA: MIT Press, 2018: 9-12.
- [9] DARLING K. Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects[M]//CALO R, FROMKIN A M, KERR I. Robot law. Cheltenham: Edward Elgar Publishing, 2016: 213-232. DOI:10.4337/9781783476732.00017.
- [10] MOYLE W, JONES C, MURFIELD J, et al. Using a therapeutic companion robot for dementia symptoms in long-term care: reflections from a cluster-RCT[J]. Aging & Mental Health, 2019, 23(3): 329-336. DOI:10.1080/13607863.2017.1421617.

三、新闻报道等网络文献

- [1] 习近平：《携手构建公正合理的全球人工智能治理体系——在 2026 世界人工智能大会暨人工智能全球治理高级别会议开幕式上的主旨讲话》，载新华网，<https://www.news.cn/politics/leaders/20260717/72728b6f94154d63b3eaaaf9808b51eb/c.html>。
- [2] 《特斯拉技术员被机器人打成重伤！向特斯拉索赔 5100 万美元》，载新浪财经，<https://finance.sina.com.cn/roll/2025-09-29/doc-infscnsr8357965.shtml>。
- [3] 《机器人在商场表演时碰倒老人，涉事商户：老人被吓到躲开时意外被碰》，载海报新闻网，<https://w.dzwww.com/sjb/index.php?a=show&c=ad&id=pddIw5TKPG3>。
- [4] 《机器人表演伤娃频发 安全疏漏亟待敲响警钟》，载闪电新闻网，<https://sdwx.iqilu.com/share/YS0yMS0xNzIzMzkzNA==.html>。
- [5] 张平：《构建安全可信可控的 AI 新生态》，载中国网信网，https://www.cac.gov.cn/2025-09/26/c_1760606717425964.htm。
- [6] 《中央网信办部署开展“清朗·未成年人网络保护”专项行动》，载新华网，<https://www.news.cn/politics/20260720/8b1d1378454848eeb3b70fc5de840180/c.html>。
- [7] 新西兰中文先驱网：《跟 AI 聊天一年 24 岁女子选择去死 竟获多次鼓励》，<https://www.chineseherald.co.nz/news/lifestyle/chatgpt-death/>。
- [8] 熊润淼：《老爹信 AI 不信当医生的女儿 预防脑梗处方找 AI “过目”后才放心服药》，载网易网，https://m.163.com/dy/article/KV73FL0E05129QAF.html?spss=adap_pc。
- [9] 汪超：《生成式人工智能的深度伪造风险与伦理规制》，载中国社会科学网，https://www.cssn.cn/skgz/bwyc/202602/t20260205_5972354.shtml。
- [10] 刘典：《筑牢具身智能产业发展的安全根基》，载《学习时报》2026 年 2 月 28 日。
- [11] 《具身智能机器人走进千家万户，难在哪儿》，载《科技日报》2026 年 7 月 2 日。
- [12] 陈博远、杨耀东：《构建全链路防御护航具身智能范式跨越与安全落地》，<https://www.secrss.com/articles/89315>。
- [13] 张凌寒：解读《人工智能拟人化互动服务管理暂行办法》，https://www.cac.gov.cn/2025-12/28/c_1768662848000498.htm?trk=article-ssr-frontend-pulse_title-text-block。
- [14] 《清华赵明国谈人形机器人的伦理与安全：有必要在行业初期制定原则》，载清华新闻网，<https://www.tsinghua.edu.cn/info/1182/113478.htm>。
- [15] 张喜玲、李军文：《数字经济视角下工业机器人对就业的影响——促进与冲击的双重检验》，载《当代经济》，<https://link.cnki.net/urlid/42.1430.F.20260701.1128.006>。

