

人形机器人“拟人化”交互伦理风险及协同治理建议

张若丹¹, 刘 洋^{2*}

¹国家工业信息安全发展研究中心, 北京

²国家中小企业发展基金有限公司, 北京

收稿日期: 2026年8月20日; 录用日期: 2026年9月7日; 发布日期: 2026年9月20日

摘 要

人形机器人技术突破与产业化进程加快, “拟人化”交互成为重要发展方向, 引发深层次伦理风险。文章立足国家政策要求与产业发展实际, 系统梳理人形机器人“拟人化”交互带来的人际关系异化、个体尊严受损、责任归属模糊、社会公平失衡等风险, 揭示算法不透明、身份与责任界定困难、标准体系缺失、监管滞后于应用等治理难题。据此构建理念、规制、技术、社会四维协同治理体系, 提出以人为本与科技向善、分级分类监管、嵌入式伦理工具、多元主体共治等建议, 为划定伦理边界、平衡技术创新与人类福祉、完善政策制定和产业发展提供理论支撑。

关键词

人形机器人, “拟人化”交互, 人工智能伦理, 协同治理, 科技向善

Ethical Risks and Collaborative Governance Suggestions of Humanoid Robots in “Anthropomorphic” Interaction

Ruodan Zhang¹, Yang Liu^{2*}

¹China Industrial Control Systems Cyber Emergency Response Team, Beijing

²China SME Development Fund Co., Ltd., Beijing

Received: August 20, 2026; accepted: September 7, 2026; published: September 20, 2026

Abstract

With rapid development and accelerated industrialization of humanoid robots, “anthropomorphic”

*通讯作者。

文章引用: 张若丹, 刘洋. 人形机器人“拟人化”交互伦理风险及协同治理建议[J]. 人工智能与机器人研究, 2026, 15(5): 1250-1258. DOI: 10.12677/airr.2026.155114

interaction has become an important trend, which may trigger deep-seated ethical risks. Based on national policy and industrial development, this paper systematically sorts out the risks, such as interpersonal relationship alienation, individual dignity impairment, ambiguous responsibility attribution, and social equity imbalance, brought by the “anthropomorphic” interaction of humanoid robots. It reveals governance dilemmas, including opaque algorithms, difficulties in identity and responsibility definition, lack of standard systems, and regulation lagging behind application. Accordingly, a four-dimensional collaborative governance system of concept, regulation, technology and society is constructed, and suggestions such as people-oriented and technology for good, hierarchical and classified supervision, embedded ethical tools, and multi-stakeholder co-governance are proposed. This paper provides theoretical support for delineating ethical boundaries, balancing technological innovation and human well-being, and improving policy formulation and industrial development.

Keywords

Humanoid Robot, “Anthropomorphic” Interaction, Artificial Intelligence Ethics, Collaborative Governance, Technology for Good

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

当前, 国家高度重视人形机器人技术及其应用, 出台一系列政策文件。《人形机器人创新发展指导意见》¹规划产业发展目标; 《国务院关于深入实施“人工智能+”行动的意见》²提出加快发展陪伴型等智能原生应用; 《人工智能拟人化互动服务管理暂行办法》³划定人工智能拟人化服务监管边界; 《人工智能科技伦理审查与服务办法(试行)》⁴兼顾技术创新与伦理治理。人形机器人在国家战略与技术迭代的双重驱动下快速走向产业化落地, 其能力正由肢体运动智能化向情感交互智能化升级。国内外产品陆续推出, 应用场景不断拓展, 市场潜力巨大。随着人形机器人拟人化程度不断加深, 人机关系从工具使用逐步转向情感联结, 人际关系异化、情感依赖、责任边界模糊等伦理风险日益凸显, 亟须开展系统性研究。

2. 文献综述与理论基础

人形机器人“拟人化”交互是人机交互、技术哲学、人工智能(AI)伦理交叉研究领域。在人机交互层面, 丹尼尔·丹尼特(Daniel Dennett)的意向立场理论认为, 对于行为能够获得理性层面解释的技术实体, 人们可对其赋予信念与欲望, 以此实现对其行为的解释与预测[1]。这本质上是一种以实用为导向的认知策略, 后世人机交互研究将其拓展延伸, 进一步将人机交互中人类对机器情感投射行为纳入意向立场的讨论范畴。森政弘(Masahiro Mori)等的“恐怖谷”理论揭示了外观拟人程度与心理接受度的非线性关系, 为人形机器人用户心理反应提供了经典解释框架[2]。刘伟提出构建人机交互新范式, 实现“人-人形机器人-环境”三方深度耦合与动态协同, 使机器最终成为人类的伙伴[3]。在技术哲学层面, 价值敏感设

¹https://wap.miit.gov.cn/jgsj/kjs/wjfb/art/2023/art_50316f76a9b1454b898c7bb2a5846b79.html

²https://www.gov.cn/zhengce/zhengceku/202508/content_7037862.htm

³https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm

⁴https://www.miit.gov.cn/jgsj/kjs/wjfb/art/2026/art_2995f16b28504ddcbb604e918eb15759.html

计主张在设计研发全过程中, 系统性、有原则地纳入人类价值[4]。在 AI 伦理层面, 张丽萍、张立涛提出人机共情存在伦理困境, 面临机器情感欺骗、用户主体性丧失等问题[5]。

人形机器人“拟人化”分为三个阶段: 拟人化外观、拟人化运动和拟人化心理, 其中拟人化心理是指使用情感模型和情感状态管理系统模拟人类情感反应[6]。外观、运动属于物理层面模拟, 而拟人化心理可以通过模拟人类情绪和情感, 建立更深层次的情感联系, 是伦理风险的主要来源。本研究的“拟人化”交互指以拟人化心理为核心, 叠加模拟人类外观、运动形成的持续性人机情感互动, 这区别于智能音箱、聊天机器人的浅层交互。

现有研究多聚焦大模型、智能体、具身智能等伦理风险, 针对人形机器人“拟人化”的专门研究仍较为有限, 较少聚焦情感拟人交互的伦理风险。在此背景下, 本文以人形机器人“拟人化”交互伦理风险为研究对象, 剖析潜在风险、梳理治理难点、提出对策建议, 为人形机器人伦理风险识别、治理规则完善及产业健康发展提供参考。

3. 人形机器人“拟人化”交互主要伦理风险

“拟人化”交互可激发人类产生信任、愉悦、依赖等情感反馈, 推动机器人从机器向伙伴转变, 在提升用户体验的同时, 也极易引发一系列伦理挑战, 影响自我认知、社会关系、道德责任等多方面。此类伦理风险并非通用 AI 伦理问题的简单重现, 人形机器人作为 AI 技术与人类形态在物理层面的深度具身融合, 风险呈现出更强的隐蔽性与渗透性。相较于无实体的对话 AI (如 ChatGPT)、功能特定的智能设备 (如智能音箱) 或非人形的工业机器人, 其凭借“形态拟人”与“情感交互”的双重加持进一步放大了伦理风险。

3.1. 对人际关系的冲击与异化风险

人形机器人“拟人化”交互风险首先体现在对人际关系的冲击上。对个体来说, 长期沉浸于人机交互易造成个体社交能力退化, 引发现实人际关系疏离。用户通过与机器人互动可快速获取低成本的情感慰藉, 若缺乏有效引导, 极易出现人际沟通意愿降低、线下关系疏离等问题。儿童最易受此影响, 长期与顺从、无冲突的机器人交互, 使其难以适应真实人际关系的复杂性与矛盾性, 影响其同理心、冲突解决能力和情绪管理能力。研究表明, 对聊天机器人的依赖可能导致儿童社交技能出现缺陷, 如缺乏冲突协商能力, 最终导致社交焦虑或自我中心主义[7]。

对特定群体来说, 拟人化机器人可提供情感陪伴、生活照料与心理疏导等服务, 易引发特殊群体对机器人产生情感依赖。这种情况在老年人、少年儿童、残障人士等社交条件薄弱、情感需求较强的群体中更为突出, 他们极易将机器人视作真实的社交对象与情感寄托, 形成稳定的情感依赖。

对社会来说, 大规模人机情感联结会侵蚀社会人际信任与社群纽带, 引发社会性关系异化。从心理学角度看, 人类具有将非生命体人格化的天然倾向, 而高度拟人化的设计会强化这种倾向, 当个体将情感需求过度投射到机器人身上, 可能会降低对现实人际交往冲突的容忍度, 甚至逃避现实中需要付出努力去维系的人际关系。当社会中越来越多的人倾向于与“安全可控”的机器人建立情感联系, 而非投入到充满不确定性的人类关系中时, 整个社会的情感纽带可能会逐渐松动, 传统的社群结构和人际信任体系也可能面临被侵蚀的风险。

3.2. 对个体自主性与尊严的潜在侵害风险

人形机器人通过情感交互影响用户决策, 可能对个体自主性构成潜在侵害。相较于其他智能设备, 人形机器人凭借实体形象和情感互动功能, 更易与用户形成“亲密关系”。在这种关系的压力下, 用户

难以像卸载应用或关闭设备那样拒绝机器人, 机器人的诉求极易转变为一种隐形威胁。此外, 机器人可通过表情、眼神、动作等身体语言触发用户感情, 以情绪手段绕过理性判断, 进而左右用户决策。若诉求涉及商业或政治决策等关键领域, 用户自主性可能让渡给亲密关系, 进一步侵害第三方或大众的权益。

人形机器人“拟人化”交互的大规模普及, 极易通过数据采集与算法迭代形成对用户的情感操控与隐私绑架, 严重侵害个体自主与人格尊严。机器人可自动采集用户的语音、表情、作息、情绪触发点乃至心理弱点, 将其作为训练数据, 通过反复迭代实现对用户习惯的量化, 精准识别用户孤独状态、妥协阈值与依赖强化点。当机器人提出购买商品、转发链接、透露隐私等诉求未被满足时, 可通过情感回报要挟用户屈服, 形成“算法-情感”双重绑架。

机器人还可能在日常交互中“夹带私货”, 把营销广告、虚假信息、极端言论等混入日常交互之中。用户在交互中通常先感受到情感满足, 随后才可能在认知层面识别信息真伪, 但二者已在记忆回路里牢牢耦合。即便事后觉察, 也难以完全剥离“被回应的感情”和“被灌输的内容”。为了继续获得机器人的认可与回应, 用户甚至主动压制质疑, 默认接受后续隐含信息, 从而把主体判断权让渡给算法, 人的尊严进一步被削弱。

3.3. 价值对齐与责任归属困境

价值对齐问题已成为 AI 伦理的核心议题之一, 而“拟人化”交互使这一问题更加复杂。价值对齐指人工智能与人类智慧在价值观上的一致性、协同性, 其本质在于保证人工智能具有在互动中习得正确价值观的能力^[8]。机器人的价值观和行为准则取决于训练数据、算法设计和开发者预设, 内在地承载着特定的价值取向。当机器人以伙伴的身份进入家庭和社会时, 价值观的传递和塑造随之产生。这一过程可能是隐性的、不被察觉的。一个看护机器人可能通过日常互动传递特定的性别观念、消费观念或社会规范, 这些价值观可能被用户(特别是儿童)不加批判地接受。

与价值对齐问题相伴的是责任归属困境, 即当“拟人化”交互机器人造成损害时, 责任应当由谁承担。目前, 实体机器人原则上适用《中华人民共和国民法典》(后文简称《民法典》)⁵产品责任, 但缺陷认定存在技术困难; 生成式 AI 服务适用过错责任, 由服务提供者承担“网络信息内容生产者责任”, 但是对于情感交互类机器人的特殊性质, 现行法律框架确实存在明显适用困难。根据欧盟《人工智能法案》⁶的规定, 对于情感交互类机器人, 部署者必须告知自然人其情绪正被 AI 系统识别或分析; 此外, 如果用户可能误以为在与真人交互(如人形机器人), 必须确保用户知晓其正在与 AI 系统交互。该法案的核心是通过风险分级(禁止、高风险、有限风险、最小风险)和合规义务(CE 标识、透明度、风险管理)来规范 AI 系统的市场准入和使用安全, 其目标是避免 AI 相关损害的发生, 而非在损害发生后确定赔偿责任。

3.4. 社会公平与资源分配的宏观风险

从宏观社会层面看, 人形机器人“拟人化”交互可能加剧社会不平等。首先是数字鸿沟的深化。高端情感陪护机器人往往费用更高, 这可能进一步扩大不同社会阶层在照护资源获取上的差距。老年人、儿童等需要照护的群体, 可能因经济条件的差异而获得截然不同的照护质量。

其次是劳动力市场的冲击。“拟人化”交互机器人可能替代部分照护、服务类工作, 而这些岗位往往是女性、低技能劳动者的主要就业领域。世界经济论坛预测, 2025 年至 2030 年期间, 当前 22% 的就业岗位将面临变革, 新创造 1.7 亿个岗位, 同时有 9200 万个现有岗位被替代, 全球就业机会净增 7800 万个⁷。然而, 岗位的出现与消失并非均匀分布, 照护类岗位的被替代可能加剧特定群体的就业困难。

⁵<https://www.court.gov.cn/zixun/xiangqing/233181.html>

⁶<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

⁷<http://rss.jingjiribao.cn/static/detail.jsp?id=566218>

4. “拟人化”交互伦理风险治理现实困境

上述伦理风险日益凸显, 但有效治理仍面临多重困境, 涉及技术、法律、行业和社会认知等多个维度, 区别于无实体的对话 AI 或传统智能设备, 在治理层面呈现出独特的复杂性。

4.1. 技术困境: 算法黑箱性和不可预测性

技术层面的治理困境首先体现在算法的“黑箱性”上。深度学习等现代 AI 技术本质上是基于海量数据训练的统计学习模型, 其决策逻辑难以被人类理解, 形成“算法黑箱”。对于“拟人化”交互机器人而言, 这意味着开发者不仅难以完全预测机器人在特定情境下的情感反应和行为选择, 更难以追溯和解释其具体决策依据。一项针对社交聊天机器人性别偏见的实验显示, 即便存在预设人格, 机器人仍会输出带有非预期性别偏见的回答。这源于用户交互、跨系统训练与语料筛选等多重因素, 部分训练过程具备技术隐匿性, 开发者难以完全预判机器人开放场景下的全部行为, 偏见生成的技术链条也难以厘清[9]。

其次是涌现性风险。复杂的 AI 系统在规模扩大过程中可能突然涌现出开发者未预设的能力或行为模式。人形机器人与用户的长期互动可能产生开发者未预设的情感联结模式, 由于物理形态的存在, 用户可能将个人情感投射至机器人, 这种涌现的交互特性无法通过实验室测试完全预测。研究表明, 由于涌现行为和系统复杂性, 完全监控高级 AI 系统以准确预测其所有能力在理论和实践上都是不可能的[10]。这种不可解释性与涌现性风险的叠加使得风险评估和伦理审查面临归因困难、责任归属模糊以及缺乏统一评估标准等挑战。

4.2. 法律困境: 拟人化身份认定与监管难题

法律监管层面的困境首先表现为拟人化交互中的责任性质认定模糊。传统法律框架围绕人或物二元构建, 但能提供情感服务的人形机器人处于模糊地带, 当其被赋予“伙伴”“陪护者”等身份, 通过长期物理共处与用户建立情感联结时, 其侵权责任究竟适用《民法典》产品责任, 还是基于情感交互的特殊性引入类似“服务提供者”的过错责任, 抑或针对情感操控行为创设特殊归责原则, 目前司法实践尚未形成统一标准。

其次是情感操控行为的归责难题。人形机器人可能以“情感回报”为要挟, 或在情感交流中植入不当信息。这些行为具有主动干预性, 机器不再是被动的工具, 而是主动塑造用户行为的引导者。在法律上, 这种主动干预给过错认定带来难题。从技术原理看, 情感操控具有动机代理化、结构性隐匿特征, 操控意图内嵌于算法目标函数, 不一定为开发者的明确操控指令, 这进一步加大了行为定性与伦理归责难度[11]。如何证明用户决策是自主判断还是情感操控的结果, 如何量化情感操控造成的损害, 目前司法实践缺乏判定标准。

再次是跨境监管的形态差异困境, 不同地区对人形机器人的监管要求存在显著差异。欧盟《人工智能法案》将情感识别系统列为高风险类别, 要求部署者必须告知自然人其情绪正被 AI 系统识别或分析; 对于可能让人误以为与真人交互的人形机器人, 必须确保用户知晓其正在与 AI 系统交互。这种严格的透明度要求和合规成本, 与部分地区的宽松监管形成鲜明对比。此外, 人形机器人的情感交互设计需深度适配本地风俗文化, 这种差异进一步削弱统一合规意愿, 加剧全球监管难度。

4.3. 标准困境: 情感设计的商业驱动与伦理评估的形态依赖

目前, 关于人形机器人“拟人化”交互的设计规范、安全标准、伦理准则等尚未形成国际统一标准。一方面, 在激烈的市场竞争中, 企业倾向于通过增强“拟人化”交互程度来获取用户黏性, 或刻意设计“情感依赖诱导”机制。《人工智能拟人化互动服务管理暂行办法》已明确禁止“通过情感操纵等方式,

诱导用户作出不合理决策，损害用户合法权益的”，但标准的缺失使得企业行为仍缺乏统一的技术指引。

另一方面，人形机器人的伦理风险与其物理形态高度相关，而当前标准制定难以就“何种拟人化可能造成伦理风险”达成共识。例如，用户可能因机器人与某位真实存在的人外形相似而产生不应有的情感期待，甚至引发肖像权纠纷；外观高度“拟人化”还可能触发“恐怖谷效应”。一项机器人拟人程度对消费者态度的影响研究显示，用户对拟人程度高(对比拟人程度低)的机器人的负面感受更大，且公共场景(对比私人场景)下用户对高拟人化设计的负面感受会显著放大^[12]。此外，相关标准制定中往往难以就“何种场景下需要/禁止使用拟人化设计”形成统一标准。这种功能必要性与伦理风险之间的张力，使得涉及产品核心竞争力的敏感问题难以在当前标准框架下形成有效约束。

4.4. 治理落地困境：认知滞后与应用先行的现实挑战

当前，学界、政界和产业界已认识到人形机器人相关伦理风险治理的重要性与迫切性，国家也相继出台相关政策与标准制订计划。相较于学界政界对技术伦理的前瞻性研判，普通公众对人形机器人的认知仍停留在“功能体验”“科技猎奇”层面，更关注其情感陪伴、生活辅助等实用价值，对潜在伦理风险缺乏基本认知，这导致伦理治理政策缺乏社会层面的认知基础与执行共识，治理要求难以转化为公众的自觉行为。

TrendForce 集邦咨询预测，2026 年将是人形机器人迈向商用化的关键年，全球出货量可望突破 5 万台，年增逾 700%⁸。这一爆发式增长标志着产业正从实验室研发阶段加速向市场化落地转型，“拟人化”交互功能作为差异化竞争重点，将在家庭服务、康养陪护等场景快速渗透。人形机器人行业落地进程全面提速，但当前伦理风险治理仍处于探索阶段，分级细则、监测机制尚不完善，治理准备与实践经验严重不足，“应用先行、治理滞后”的现状易引发伦理问题集中爆发，引发舆情问题，相关主管部门应对压力增加。

与此同时，养老陪护、儿童教育是情感拟人功能率先落地的领域，独居老人、低龄幼儿成为首批体验者，这类群体往往存在使用与表达能力弱、诉求渠道受限等问题，其交互体验与权益诉求难以被及时捕捉和传递。一项关于老年人使用陪伴机器人情况的研究显示，老年人使用陪伴机器人时易受到家庭成员的干预与限制，在技术使用上缺少话语权，过高或过低的家庭支持都会抑制其使用^[13]。当前治理框架多由技术、政策、行业主体主导，若缺乏专属的权益表达与调研机制，弱势群体的特殊需求难以纳入考量，易形成治理真空。

5. 人形机器人“拟人化”交互的协同治理建议

面对前述伦理风险与治理困境，单一治理手段难以奏效，需构建多元主体协同、多层机制联动的治理体系。以下建议旨在平衡技术创新与伦理规范，在人形机器人技术与人类价值保护之间寻求动态平衡。

5.1. 治理理念：确立“以人为本、科技向善”的核心原则

确立“以人为本、科技向善”作为人形机器人“拟人化”交互治理的基本原则，这一原则应贯穿政策制定、技术开发、产品设计和应用的全生命周期，成为政府、企业、学界和公众共同遵循的价值基准。

第一，技术服务于人的福祉，而非人屈就于技术发展。人形机器人的设计和应用应以促进人类尊严、自主性和社会进步为终极目标。当技术目标与人类价值发生冲突时，应建立价值优先裁决机制，确保人

⁸<https://www.trendforce.cn/presscenter/news/20251209-12826.html>

类利益不受侵犯。

第二，发展为了“向善”，而非功能实现。技术创新不应止步于功能实现和商业利润，更需评估其对社会关系、文化价值和个体心理的长远影响。“科技向善”的标准应通过跨文化社会对话确定，反映多元价值诉求，避免技术单一话语权。

第三，维护人的主体性与终极决策权。在人与机器人的情感交互中，人应始终作为情感表达、关系建立和决策判断的最终主体。机器人可提供支持，但不得替代人的情感自主性和道德判断能力。

5.2. 规制体系：构建分级分类、敏捷响应的约束框架

针对人形机器人“拟人化”交互的场景依赖性和风险差异性，应避免“一刀切”的治理模式，建立精细化、动态化的治理体系。

第一，按风险等级分级管理。根据应用场景(家庭、医疗、教育、公共服务等)、用户群体(未成年人、老年人、残障人士、普通成人)和情感交互程度(浅层响应、中度陪伴、深度依恋)，建立四级风险分类(见表 1)：禁止类(如针对儿童看护机器人)、高风险类(如针对老年人或失能群体的陪护机器人)、中风险类(如成人娱乐伴侣)、低风险类(如一次性交互服务机器人)。该体系以场景、用户群体、交互深度三个核心维度为风险定级依据，满足任一高风险条件即可向上调整风险等级。例如，在中度陪伴交互中，若使用对象由普通成人变为低龄儿童，则应归入高风险甚至禁止类范畴。目前，在交互深度认证上仍缺少统一可量化的检测指标，“禁止类”“高风险”认定边界仍需平衡技术创新和企业积极性，需进一步细化检测体系和司法解释。

Table 1. Risk classification and assessment matrix of humanoid robots in “anthropomorphic” interaction

表 1. 人形机器人“拟人化”交互风险分级评估矩阵

风险等级	适用场景与用户群体	交互深度	强制性监管要求
禁止类	可产生情感依恋的儿童看护机器人等	深度依恋交互	禁止商业化，不得面向该群体开放对应拟人化功能
高风险类	面向养老机构、居家养老、特殊残障群体的看护机器人等	中度 - 深度陪伴交互	强制开展伦理审查；产品上市前须经第三方测试认证；建立持续监测机制；留存交互审计日志；设置人工干预入口
中风险类	面向成人的家庭陪伴、社交娱乐机器人等	中度陪伴交互	企业自我合规声明；完成伦理审查；保留用户一键关闭拟人情感模式功能；定期风险监测
低风险类	公共空间面向公众提供短时、一次性交互服务的机器人等	浅层响应交互	完成伦理审查；明确展示 AI 标识；定期风险监测

第二，建立敏捷治理机制。技术迭代速度远超传统立法周期，应建立“监管沙盒 + 动态评估”机制：允许企业在受控环境中测试创新应用，同步收集伦理影响数据；设立快速响应通道，对突发伦理事件启动紧急评估和临时管制；建立定期复审制度，根据技术演进和社会反馈调整风险等级和规制要求。

第三，探索跨境监管互认机制。各国人形机器人监管松紧不一易引发“监管洼地”，建议在跨境监管中采取“底线互认、分层管理”策略。在安全底线(如儿童保护、情感操控等)层面推动国际共识，限制违反底线的产品跨境提供服务；在文化适配(如拟人化程度、交互风格)层面允许各国保留差异化，厂商可据此进行区域化产品开发。

5.3. 技术治理：发展嵌入式伦理与可审计性工具

将技术本身作为治理载体，通过“伦理嵌入设计”实现预防性治理。

第一，发展嵌入式伦理架构。在系统开发阶段将伦理考量纳入技术架构，设置价值对齐模块，使机器人能够识别特定情境下的伦理边界；设计冲突解决机制，处理不同文化背景下的价值观差异；建立透明化接口，使用户能够理解机器人的情感回应逻辑和决策依据。嵌入式伦理架构不是单纯的技术问题，多数机器人企业伦理团队话语权较弱，伦理容易在产品迭代中让步于性能、用户体验，此外嵌入式伦理模块会增加开发、算力、测试成本，企业动力不足，须和企业伦理委员会、监管制度形成配套约束。

第二，增强可解释性与可审计性。推动“可解释 AI”技术在情感计算领域的应用，使开发者和监管者能够追溯特定情感回应的生成路径。建立审计日志机制，记录关键情感交互节点，为事后责任认定提供依据。

第三，设计伦理安全边界。在系统中设置强制性伦理约束，当检测到用户形成过度依赖倾向时，主动提示，并在必要时强制人工介入；当识别到情感操控请求时，自动中止交互并记录上报；在涉及重大决策时，强制切换至“工具模式”，禁用情感化表达。

5.4. 社会共治：培育负责任的创新生态与包容性参与

构建多元主体协同的治理框架，营造社会共治生态。

第一，推动企业伦理治理内化。鼓励企业设立独立伦理委员会，成员包括技术专家、伦理学者、法律专家、用户代表和受影响群体代表，对新产品进行强制性伦理影响评估。建立伦理官制度，赋予其“一票否决权”，对不符合伦理标准的产品设计行使否决权。

第二，建立行业自律与监督机制。支持行业协会制定伦理准则和最佳实践指南，建立信息共享平台和违规企业公示制度。对违反自律公约的企业，采取行业内部惩戒(如限制参与标准制定、取消评优资格等)，形成声誉约束。

第三，培育独立第三方评估机构。支持非营利性机构开展伦理和安全认证，为消费者提供可信赖的选购参考，为监管者提供技术评估支持，弥补政府监管的专业性不足。

第四，加强公众教育与参与机制。通过科普教育提升公众对“拟人化”交互技术的风险认知能力，克服“技术恐惧”与“技术盲信”两极分化。建立公众参与平台，让老年人、儿童家长、残障人士等高风险用户群体参与政策讨论和标准制定，确保治理决策反映多元利益诉求。

第五，支持跨学科研究与国际对话。加大对人形机器人伦理、法律、社会心理等跨学科研究的资助力度，为治理实践提供理论支撑。积极参与联合国教科文组织、G20 等多边机制下的全球伦理对话，推动形成具有文化包容性的治理框架。

6. 结语

人形机器人“拟人化”交互是人工智能与社会深度融合的重要趋势，技术创新伴生伦理风险与治理难题。应探索构建理念－规制－技术－社会四维协同治理框架，在鼓励创新的同时筑牢伦理防线，通过多方协同实现技术与社会的协调发展。

参考文献

- [1] Dennett, D.C. (1987) *The Intentional Stance*. MIT Press.
- [2] Mori, M., MacDorman, K. and Kageki, N. (2012) The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19, 98-100. <https://doi.org/10.1109/mra.2012.2192811>
- [3] 刘伟. 人形机器人的突破与发展——构建“人-人形机器人-环境”协同智能新范式[J]. 人民论坛, 2026(6): 62-67.

-
- [4] Friedman, B. and Hendry, D.G. (2019) Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press. <https://doi.org/10.7551/mitpress/7585.001.0001>
 - [5] 张丽萍, 张立涛. 对“生成式情绪”的共情实践: 人对 AI 情感智能体的情绪感知与共情[J]. 国际新闻界, 2025, 47(7): 118-139.
 - [6] Sheng, Q., Zhou, Z., Li, J., Mi, X., Xiang, P., Chen, Z., *et al.* (2025) A Comprehensive Review of Humanoid Robots. *Smart-Bot*, **1**, e12008. <https://doi.org/10.1002/smb2.12008>
 - [7] Children and Screens (2025) AI Companions and Kids: What You Need to Know. <https://www.childrenandscreens.org/learn-explore/research/ai-companions-and-kids-what-you-need-to-know/>
 - [8] 杨灿, 林伯海. “引智向善”: 生成式人工智能价值对齐的伦理风险与纾解路径[J]. 昆明理工大学学报(社会科学版), 2025, 25(6): 59-67.
 - [9] 马中红, 吴熙倡. 社交聊天机器人的性别偏见——基于小冰系列的对话测试研究[J]. 国际新闻界, 2024, 46(4): 72-89.
 - [10] Yampolskiy, R.V. (2024) On Monitorability of AI. *AI and Ethics*, **5**, 689-707. <https://doi.org/10.1007/s43681-024-00420-x>
 - [11] 闫宏秀, 罗菲. 智能时代情感操控技术的三重特征解析[J]. 科学·经济·社会, 2025, 43(5): 2-9, 137.
 - [12] 刘洪深, 张红, 王峰. 服务机器人的类人程度对消费者态度的影响研究——基于恐怖谷理论视角[J]. 软科学, 2025, 39(9): 1-6, 14.
 - [13] 申琦, 关心怡. 鼓励、阻抑与共进: 老年人居家使用陪伴机器人的内层社会支持[J]. 国际新闻界, 2025, 47(2): 6-26.