工程与管理科学
Engineering and Management Science
第 7 卷◆第 7 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

具身智能机器人在博物馆应用落地的探索

朱凌梅
中国国家博物馆　100016

DOI:10.12238/ems.v7i7.14268

[摘　要] 随着大语言模型驱动的具身智能技术日趋成熟，其在博物馆领域的创新应用正逐步成为现实，预示着文化传播与场馆教育模式的重大变革。将具身智能机器人引入博物馆，也伴随着一系列不容忽视的风险与挑战。本文对此进行了系统性梳理，着重探讨了信息与数据安全、人机物理及认知安全、观众过度信任引发的社会伦理风险，以及技术自身带来的潜在影响。基于此，文章结合国标要求提出了详尽的风险应对策略。本文旨在为博物馆安全、负责任地部署具身智能提供理论支撑与实践参考，共同探索在保障观众安全与体验的前提下，具身智能赋能文化遗产保护与教育的广阔前景。

[关键词] 具身智能；博物馆；风险管理；数据安全；人机交互；文化传播

一、大模型与具身智能

近年来，DeepSeek、GPT、Gemini、Claude 等大语言模型在自然语言理解、对话生成和任务规划等方面取得突破，并通过多模态接口逐步拓展至图像识别、语音合成、动作生成等领域。这些能力为具身智能系统提供了“认知层”的支持，使其不仅能听懂指令，还能基于环境感知做出符合语境的响应。

具身智能（Embodied intelligence，EI）指的是机器人等智能体通过其物理形态与外部环境之间的紧密互动和反馈，来感知周围世界、做出决策并执行相应动作的能力[1][2]。智能体不仅能感知外部环境，还能通过与环境的实时互动，利用深度学习与强化学习等模型，基于物理经验不断学习和适应[3]。

图 1　Google DeepMind 的机器人具身大模型 “RT-2”

（图片来源 Google DeepMind 主页）

目前，具身智能如 Pepper 机器人在博物馆中的应用多为试点项目、研究性质或作为辅助角色，旨在探索机器人如何提升参观体验。

二、具身智能机器人在博物馆等公共服务场所落地时可能面临的风险和问题

在互联网广为流传的一个视频中，一台机器人开机后，开启姿态稳定程序而在悬挂状态下试图调整自身平衡，因控制发散持续挥舞肢体，打翻桌子与电脑[4]。这对具身智能机

器人在公共空间的使用而言具有重要警示意义：在密集人群、高价值展陈、儿童频繁出入的博物馆、商场、医院等环境中，部署具身智能机器人应预先评估其可能带来的风险及问题，主要包括物理碰撞与误伤、功能失效或偏移、信息安全和数据安全风险等。安全是人工智能落地应用时应该关注的核心问题[5]。

1.物理碰撞与误伤

在博物馆等公共空间中，可移动的服务类机器人因其自重原因具有较大惯性，其任何肢体的非预期运动都可能造成实际伤害或引发恐慌。比如，行进中碰撞或碾压人体、因姿态不稳碰撞周围观众或设施、在实施指引动作时延展肢体接触人或展品等。目前的机器人的在防碰撞方面主要依赖自身运动反馈机制，缺乏感知自身输入可靠性的能力。这对具身智能在公共场所中的感知和判断能力、面向失配场景自我限制的主动约束能力，以及在异常情况下的响应速度提出了较高的要求。

2.功能失效或功能偏移

在博物馆等公共空间中，具身智能可能会受到环境、观众行为、幻觉或机械结构影响导致功能失效或偏移。

1）受外部物理环境干扰导致功能偏移

具身智能机器人在博物馆等公共空间工作时，当人群聚集、布展环境改变或者空间内光线有较大变化和干扰时，可能会因感知失真、决策逻辑冲突造成功能失效，例如行进路线规划错误、外部请求误判、障碍物误判、误触人体或展品等。

2）人机交互不确定性引发的失效

人机交互过程具有多维度、多层次特点。主要有用户表达差异或语义模糊、故意干扰等情况。用户发送的请求可能因使用方言或俚语，或表达不够清晰而被具身智能机器人误读，造成无响应或错误响应；错误的肢体动作响应又可能引发误解或带来安全隐患。儿童或青少年观众可能出于好奇，通过遮挡传感器、故意拉扯、发送异常指令或通过 Prompt 注入等方式对具身智能机器人进行干扰。

3）因产生“幻觉”现象而造成功能失效风险

基于机器学习的 AI，尤其是使用神经网络、深度学习训练的模型等，几乎必然会有“幻觉”。大语言模型可能会因为“幻觉”而编造知识、历史事件、评价等内容；听觉、视觉模型可能会因“幻觉”识别出不存在的声音或影像[6]。这些错误的输出、输入可能会令具身智能的功能失效，带来不良体验或造成恐慌。

3.信息安全及数据安全风险

具身智能在对博物馆观众进行服务时需要进行多模态数据采集，并且在完成知识整合后按要求进行输出。采集的数据中可能包括：观众个人信息（可能包括未成年人的面部图像）、场馆的环境信息和位置信息、藏品和展览信息等。这些信息如果处置或保管不当，可能会引发法律纠纷[5][6]、危害财产安全或触犯数据安全法。这些数据聚合起来，可能构成比简单的面部识别更深层次的隐私侵犯。此外，具身智能也可能因遭到网络入侵，进而发生信息窃取、传感器非法调用、输出非法内容等事故。

4.引发观众过度信任风险

由于具身智能采用了高度仿真和自然的交互设计，能够准确理解环境和观众需求，并用人性化且符合情境的语言和行为方式对观众进行输出和反馈，使得观众容易误以为其具备全面的认知和情感能力，进而在其职责范围之外寻求帮助，或产生不良互动。这种错认不仅可能影响观众的体验和情绪，还可能因误导观众造成事故和伤害。

三、风险防范与应对措施

1.标准框架与适配基线

要有效应对这些挑战，首先应确保具身智能的设计符合 GB/T 36530-2018/ISO 13482：2014《机器人与机器人装备 个人助理机器人的安全要求》。该标准为近距离与人类互动的移动机器人提供了一套系统性的安全设计与使用规范，规定了以“防止对人造成危害”的基本安全目标，涵盖了多个安全维度要求，是所有机器人投入实际使用前必须满足的基础框架。标准要求开发者在设计和部署阶段系统性地分析所有可能的危险源并进行风险评估和缓解措施设计，在将具身智能系统引入博物馆环境时，应被作为基础性原则纳入项目设计的初始阶段，并在全生命周期中始终予以参照。

此外，具身智能的多模态感知应结合场景动态更新，任何技术部署必须遵循“（人身和文化遗产）安全＞游客体验＞

效率提升”的优先级原则。

2.应对物理碰撞或误伤

在防范具身智能机器人在公共场所可能引发的物理碰撞与误伤风险方面，在充分遵守需从机械外形设计、自动控制与 AI 智能控制三个层面进行一体化考量与优化。

在机械外形设计上，应充分引入“柔性和安全”理念来设计具身智能机器人外部结构。例如，整体上应保证圆滑和重心稳定；对于需要通过肢体延展来完成的功能，如指引等，考虑以光束、声音指引代替；对于必须具备的实体肢体，采用软性、弹性材料制作，并做好动作的物理限幅。

在自动控制方面，应考虑采用高精度运动控制系统与实时反馈机制，确保机器人动作精确受控；同时，对动作执行过程中的碰撞反馈和异常反馈，应采用控制算法及时修正甚至中止操作。尤其是双足机器人，应充分考虑力反馈对安全交互的重要作用[7]和“柔顺控制（compliance control）—抵抗扰动”策略[8]。

在 AI 智能控制层,核心在于感知与决策层的前瞻性与稳健性。应引入多模态感知融合机制，实现对周围人群、展品与障碍物的高精度实时感知并对可能出现的交互、移动、聚集进行趋势判断。同时，应确保安全策略优先于任务完成。

在这三个层面的控制优先级上，机械设计是基础保障，在实时运行状态中实时响应异常反馈，不需要等待 AI 评估；而 AI 智能控制则负责“情境理解与决策前置”，判断当前环境是否适合执行某项任务、是否需要降低动作幅度、甚至是否暂停服务。越靠近事故风险的当下时刻，物理与自动控制优先级越高；越靠近任务调度和人群理解的上游阶段，AI 智能控制优先级越高。

3.应对功能失效或偏移的策略和解决方法

尽管具身智能具备强大的多模态处理能力，但在复杂的博物馆场景中，仍可能面临由环境因素、交互不确定性、模型幻觉和机械误差引发的功能偏移或失效。基于上述分析，建议以通用大模型为基础,采取“语境注入 + 行为约束 + 本地知识增强”的策略，从以下四个方向设计应对机制：

1）应对“环境干扰”引发的感知偏移与路径错误

主要从丰富数据角度出发，及时更新本地空间地图、为模型提供典型光线条件、典型噪声条件下的场馆内图像。此外，应采取行为限定策略，对于特定空间区域、人流高峰期时段，设定保守行为参数（如减速、频繁重新识别等）。

2）应对人机交互不确定性引发的失效

应对方言、俚语方面，主要从丰富语料库和设计容错措施的角度进行考虑：尽量采用专用的语音识别模型，也可考虑在多次识别失败时引导观众通过文本输入或提供可能的文本选项完成提问。应对观众的不规范和模糊提问，可采用提示模版引导用户规范表达，收敛问题，提升识别准确率。

为提升系统对恶意干扰行为（如遮挡、拉拽等）的识别能力，可在模型训练阶段加入相关行为样本，或在部署阶段通过融合视觉与触觉传感器设定阈值和响应规则，以实现异常检测并触发保护模式。

为防止非授权用户通过输入或对话引导具身智能进入不当角色或执行非预期任务，应通过权限设定限制关键指令的触发范围，防范“prompt 注入”等滥用行为，确保系统运行在可控范围内。

3）应对模型“幻觉”造成的功能失效风险

应对语言大模型或多模态模型生成内容时可能产生“幻觉”的风险，应通过限定识别范围、任务边界设置和多模态交叉验证等方式，限制其自由生成或误识的空间，以降低幻觉风险并确保行为稳定性。例如要求大语言模型响应时必须以检索到的知识库内相关文档作为依据，限制视觉模型只识别少数特定类别物体，使用特定呼唤词激活语音生成模型等。

总之，在博物馆场景下，应对具身智能系统功能偏移与失效的有效路径，并非简单依赖更强的模型能力，而是按照 GB/T 36530-2018/ISO 13482：2014 中强调的“风险可控”、“行为可预测”原则，通过任务限定、场景知识增强、行为约束与协同机制，形成“可控的智能”。

4.对于信息安全及数据安全风险的防范

应对具身智能机器人可能带来的信息安全及数据安全风险，除按照相关国家标准严格做好数据采集、传输、存储、应用过程中的安全防护措施外，还应该针对具身智能本身的物理性、高交互性和学习能力所带来的独特风险[6][9]进行下列

特殊考虑：

1）数据“去语境化”处理

对采集的原始环境数据进行脱敏处理（如去语境化），确保无法反向推导出特定个体的具体行为或身份。

2）物理隐私区域设定：划定机器人不应进入或不应进行高精度感知的“隐私区域”，例如洗手间入口、休息区等。

3)交互行为记录的精细化控制:明确记录哪些交互行为，只记录满足服务必要性的信息，避免过度记录闲聊或非服务目的的对话内容。

4)物理安全与网络安全的融合审计:将机器人硬件安全、固件安全与网络安全视为一个整体进行风险评估。例如，审计机器人是否有未授权的 USB 接口、易被篡改的内部组件等。

5） 远程急停和物理锁定机制：确保在遭受网络攻击或检测到物理威胁时，可以远程（甚至手动）对机器人进行强制停机或锁定，切断其物理活动能力和数据传输。

5.观众过度信任风险的应对措施

应通过技术限制、内容设计、人机协作管理和用户教育等多维度手段，以透明化、职责边界清晰化、引导合理预期、安全教育先行为核心原则，有效地管理和降低观众对具身智能的过度信任风险，确保博物馆服务的安全、顺畅和愉悦。此外，博物馆工作人员应充分了解具身智能的功能、局限性和潜在的信任风险，并接受相应的培训[10]。

四、总结与展望

具身智能的融入，无疑为博物馆的文化传播与场馆教育带来了前所未有的机遇。然而，要充分发挥其潜力，并避免可能对观众造成负面影响，则必须以全面而细致的风险管理为前提。

本文所构建的风险应对策略，在充分借鉴现有机器人安全标准的基础上，尤其关注具身智能在博物馆这一高人流、深文化场域中，如何有效处理观众的认知差异、语言理解、以及人机之间信任感的建立与维护等复杂挑战。只有在保障全面安全与合规的前提下，才能真正开启具身智能在博物馆中的广阔应用前景，这需要技术研发者、博物馆从业者与观众的共同努力与持续探索。

[参考文献]

[1]URING A M.Computing machinery and intelligence[M].New York：Springer，2009.

[2]FAN H，LIU X，FUH J Y H，et al.Embodied intelligence in manufacturing：Leveraging large language models for autonomous industrial robotics[J].Journal of Intelligent Manufacturing，2024：1-17.

[3]陶永，万嘉昊，王田苗，等.构建具身智能新范式：人形机器人技术现状及发展趋势综述[J/OL].机械工程学报，2025[2025-05-08].

[4]察里津的索索-FERS.智械危机！国内一厂商在调试机器人时发生失误，机器人暴走并袭击工程师[EB/OL].（2025-05-03）[2025-06-10].https：//www.bilibili.com/video/BV1ouV5zSENA/?spm_id_from=333.337.search-card.all.click

[5]吴汉东.人工智能时代的制度安排与法律规制[J].法律科学（西北政法大学学报），2017（5）：128-136.

[6]赵月，何锦雯，朱申辰，等.大语言模型安全现状与挑战[J].计算机科学，2024，51（1）：68-71.

[7]LI G X，LI Z J，KAN Z.Assimilation Control of a Robotic Exoskeleton for Physical Human-Robot Interaction[J].IEEE Robotics and Automation Letters，2022，7（2）：2977-2984.

[8]HUANG Q，DONG C C，YU Z G，et al.Resistant Compliance Control for Biped Robot Inspired by Humanlike Behavior[J].IEEE Robotics and Automation Letters，2022，7（5）：3463-3470.

[9]张凌寒.风险防范下算法的监管路径研究[J].交大法学，2018（4）：49-62.

[10]曹建峰.人工智能:机器歧视及应对之策[J].信息安全与通信保密，2016（12）：15-19.

作者简介：朱凌梅，1982，女，汉族，籍贯：北京，学位：硕士，职位：数据管理工程师，职称：中级，研究方向：云计算，数据管理，数据分析，数字化展览展示。