面向具身智能的系统安全研究

Research on System Security for Embodied Intelligence

贾雅清[1],徐　雷[1],陶　冶[1],王　芃[2](1. 中国联通研究院,北京 100048;2. 中国联合网络通信集团有限公司,北京 100033)
Jia Yaqing[1],Xu Lei[1],Tao Ye[1],Wang Peng[2](1. China Unicom Research Institute,Beijing 100048,China;2. China United Network Communications Group Co.,Ltd.,Beijing 100033,China)

摘　要:
随着具身智能在服务机器人、自动驾驶等关键场景的应用扩大,其安全挑战已从算法鲁棒性延伸至行为可控性、交互社会契合性及跨文化伦理适应性。分析了具身智能在感知—决策—执行闭环中的风险传播机制,探讨了策略建模、异常检测、行为可解释性及价值对齐等关键防御路径,并指出跨模态攻击、平台异构与意识形态适配等实际部署难题。提出具身智能安全体系需从"技术可控"向"系统协同"与"社会对齐"转变,构建兼具鲁棒性、适应性和文化敏感度的综合体系,是实现可信部署的关键所在。

关键词:
具身智能;系统安全;多模态鲁棒性;策略可解释
doi:10.12045/j.issn.1007-3043.2025.07.007
文章编号:1007-3043(2025)07-0041-05
中图分类号:TP18
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
As embodied AI systems are increasingly deployed in critical domains such as service robotics and autonomous driving, its safety challenges have evolved beyond algorithmic robustness to encompass behavioral controllability, social compliance in interactions, and cross-cultural ethical adaptability. It analyzes risk propagation mechanisms within the perception—decision—action loop of embodied intelligence, examines key mitigation strategies including policy robustness modeling, anomaly detection, behavioral interpretability, and value alignment. It further identifies practical deployment challenges such as cross-modal attacks, platform heterogeneity, and ideological adaptation. It proposes that the security paradigm for embodied AI must shift from "technical controllability" toward "system coordination" and "social alignment", emphasizing that constructing integrated frameworks with robustness, adaptability, and cultural sensitivity constitutes the cornerstone of trustworthy deployment.

Keywords:
Embodied intelligence;System safety;Multimodal robustness;Policy interpretability

引用格式:贾雅清,徐雷,陶冶,等. 面向具身智能的系统安全研究[J]. 邮电设计技术,2025(7):41-45.

0 引言

人工智能正从感知与语言向行动与交互拓展,而具身智能[1](Embodied AI)成为了其前沿方向。具身智能强调智能体需具备"行动"与"适应"能力,融合认知建模、机器人执行、语义理解与环境交互,使智能体能在现实世界中理解任务、决策并执行行为。这类智能体已被广泛应用于服务机器人[2]、医疗辅助[3]等领域,正深度介入社会生活。

然而,正是由于具身智能深度嵌入物理与社会环境,与传统AI相比,其安全问题也更复杂、更具现实威胁。具身智能不仅要应对数据安全、模型鲁棒性等传统问题,还面临着物理行为不可控、任务意图偏离、人机协作伦理冲突等直接且不可逆的风险。一旦具身系统发生故障或被操控,后果将不仅是系统崩溃或信息泄露,还可能涉及财产损失甚至人员伤害。

这一转变对AI安全研究提出了全新挑战。过去,AI安全多聚焦于算法模型本身的鲁棒性、对抗攻击与

收稿日期:2025-06-05

隐私保护(“模型安全”[4]、“数据安全”[5]),而具身智能的安全则是涵盖了感知、认知、决策到执行的全链条系统级问题。这带来多维挑战:如何确保感知在光照变化或欺骗下的准确性,如何防止语言模型被模糊提示误导,如何保障决策在动态环境中避免意外操作,如何设计可解释、可干预的执行架构。此外,具身智能的交互场景也使其安全问题更具社会属性和伦理敏感性(如家庭、医院场景需兼顾隐私、文化习惯与价值规范)。

在此背景下,构建具身智能安全框架不仅是“技术增强”任务,更是技术、伦理、规范的系统工程[6]。亟需一种新的安全认知视角:关注算法可控性、系统行为可预期性以及人机关系可协调性[7~8]。本文以此为出发点,梳理了具身智能在实际部署中的安全威胁、防御机制与评估挑战,为理论与工程提供结构化支持。

具体而言,本文系统梳理了具身智能的理论基础(问题要素与逻辑)、主要防护路径与机制设计、真实场景部署挑战以及未来研究方向(跨模态鲁棒性建模、社会对齐机制、跨平台评测标准等),旨在为安全可控具身智能的发展奠定基础。

1 理论基础与问题定义

具身智能系统作为认知与行动深度融合的智能体,其安全性问题具有明显区别于传统AI系统的内在机制与外在表现。安全失效在该类系统中并不仅意味着预测错误或模型崩溃,而常常直接投射为物理世界中的不可逆损害,例如误抓取、碰撞、路径偏离或违反伦理的行为输出,具身智能系统安全威胁与影响层级如表1所示。

从感知到执行的连续闭环结构构成了具身系统的基本运行单元。与传统AI模型将输入视作离散事件不同,具身智能以时间连续的状态流为基础,依赖感知模块对环境状态s的捕捉与编码,再经策略模块决策出动作a,最终由执行器将动作映射为真实世界中的物理影响。当感知输入受到干扰(如遮挡、对抗扰动或环境突变)时,其传播路径并非止步于识别误差,而将进一步扭曲整个策略函数$\pi(S)$,诱发结构性偏差。这种“误差放大—行为失控”的耦合机制,是具身系统安全问题的系统性根源。

表1 具身智能系统的典型安全威胁与影响层级

威胁类型	攻击方式示例	影响模块	后果概述
感知欺骗	对抗图像、遮挡、传感干扰	状态识别	环境理解错误,导致执行偏差
策略诱导	奖励黑客、提示注入	决策策略	行为输出异常,稳定性下降
执行扰动	控制信号劫持、碰撞操控	动作控制	硬件故障、物理伤害
交互误用	模糊指令、价值误对齐	人机协同	行为不当、信任缺失
多模态组合攻击	图像+语言联合误导	全系统链路	系统失控、链式风险传播

具身智能中的信息融合本质上是多模态与跨域的。系统需同时处理图像、文本、触觉、声音、动作等异构数据流,每种模态具有不同的噪声模型与语义结构[9]。这一特征导致安全风险具有跨模态传播性。例如,视觉攻击不仅可能破坏物体识别,也可能通过错误感知影响语言理解模块的任务解释;语言输入的不确定性则可能在动作控制中表现为轨迹不稳或区域侵入。这种模态交叉干扰显著提高了风险空间的维度,传统对单模态建模与评估的方式难以有效应对。

具身系统往往面向真实社会展开任务,其决策不止是对物理状态的映射,还涉及对人类语言社会行为规范与伦理偏好的理解与遵守。这种语境嵌入式认知任务使系统不仅要“做对”,更要“做得合理”[10]。若缺乏对“行为语境图谱”的建模,系统即便感知与执行准确,也难以保证其行为具有社会适当性。形式化而言,具身系统的行为输出可表示为如下受扰决策路径:

$$a' = \pi\left(s + \delta_p\right) + \delta_\pi \tag{1}$$

其中,当前系统状态为s,感知扰动为δ_p,策略扰动为δ_π。若期望系统输出在安全行为范围内,可设立阈值ε,当$\left\|\delta_a\right\| \leqslant \varepsilon$时,系统仍可视为稳定。该不等式构成了具身系统的基本鲁棒性约束条件。

2 关键安全机制与防御技术

在具身智能系统中,安全问题并非仅依赖某个单一模块的防护能力,而是贯穿感知、认知、决策与行为的系统性挑战。当前的研究趋势正从孤立的技术点防御转向多模态融合与全链条安全建构,强调感知鲁棒性、策略稳定性、行为可控性与语言输入的语义一致性等方面的协同机制。为了有效应对实际部署中可能遇到的攻击与异常行为,具身智能的安全机制日益体现出结构冗余、语义感知与行为反馈一体化的演进特征。例如,设各模态经过编码后得到的特征向量

分别为$v_1,v_2,\cdots,v_n$,则系统可定义如式(2)所示的对齐损失函数,用于控制多模态语义一致性。

$$\mathcal{L}_{\text{align}} = \sum_{i \neq j} \left\| v_i - v_j \right\|^2 \tag{2}$$

当该损失超过设定阈值τ时,系统判定当前存在模态冲突或欺骗输入,进而激活感知校验机制或执行安全中止。在此基础上,还可引入基于贝叶斯融合或注意力机制的模态加权模块,实现信息来源可信度的动态调整,从而在输入部分遭受局部扰动时自动切换为更鲁棒的通道组合。

进入策略生成阶段,具身系统需要在不确定环境下保持长期稳定性。深度强化学习(DRL)策略模型虽然具备一定的泛化能力,但在遭遇感知输入扰动或语言提示偏移时,往往会出现非线性策略跳变甚至安全边界外行为输出。对此,学界提出了一类对抗鲁棒性训练方法,其核心思想是通过引入最坏扰动条件,优化策略在极端输入场景下的表现稳定性。策略层面承载着“认知—行动映射”的核心功能。尽管深度强化学习与语言条件行为生成等模型具备较强的泛化能力,但同样易受到奖励陷阱、提示注入与策略漂移的影响。为此,本研究引入对抗鲁棒性训练框架[11],即在训练阶段考虑扰动最不利场景,对如下目标函数进行优化:

$$\max_{\theta} \min_{\delta \in S} E_{s \sim D}\left[R\left(s,\pi_\theta(s+\delta)\right)\right] \tag{3}$$

其中,π_θ为策略网络,δ为扰动空间中的变量,R为任务回报函数,D为状态分布。该目标函数可有效抑制策略因小幅感知变化而产生的剧烈行为波动。

在行为控制端,具身系统与传统AI的最大不同在于其输出具有可感知的现实物理后果。因此,行为可控性机制是安全设计不可或缺的一环。研究者普遍在机器人平台中引入了末端力限制器、速度范围监控、空间安全边界设定与动态缓冲区调整等模块,使系统在动作生成过程中具备即时响应与预防失控的能力。此外,为防止行为不可预测性带来的信任断裂,越来越多系统嵌入了任务解释接口,通过因果路径回放、策略意图可视化与实时用户可质询机制,增强人机共识水平[12]。与此同时,为防止策略模型在未知状态空间内产生高风险动作,可进一步引入软约束机制,通过惩罚不安全动作空间的输出概率,构建如下行为限制正则项:

$$\mathcal{L}_{\text{safe}} = \lambda \cdot E_s\left[\sum_{A_{\text{unsafe}}} \pi_\theta(a|s) \right] \tag{4}$$

其中,A_{unsafe}表示系统预定义的不安全动作集合,λ为安全惩罚系数,通过控制其权重,可动态调节策略生成中的安全保守性。

具身系统的输出最终表现为物理行为,因此执行阶段的安全性直接关系到设备安全与用户人身安全。在此阶段,系统需具备精准的行为控制与物理反馈能力。典型方法包括构建空间边界限制、轨迹可行性预测与实时控制信号监测等。研究中常引入力/扭矩传感器、速度限制器与动态缓冲区设定,以形成具备即时中止能力的动作管理机制。更进一步,为了增强系统可解释性与行为追溯能力,研究者提出将行为决策过程表示为因果链条,策略输出不仅取决于当前状态,还与历史行为路径相关。例如,将系统动作表达为:

$$a_t = f\left(s_t,h_t\right),\ h_t = \sum_{i=1}^{t-1} \psi\left(s_i,a_i\right) \tag{5}$$

其中,h_t为历史行为嵌入,ψ为嵌入函数。通过构建该因果行为图谱,系统能够支持用户对当前动作的可视化解释与路径追溯,增强用户对系统行为的理解与干预能力。

此外,随着大型语言模型在具身智能系统中嵌入程度的加深,自然语言指令成为控制决策的重要接口。然而,语言本身的模糊性、上下文依赖性与开放性,极易成为提示注入攻击与语义诱导的入口。针对这一问题,研究者在系统中逐步引入语义一致性判别机制。例如,通过计算视觉感知结果与语言输入之间的语义匹配度,当匹配度低于阈值时自动触发确认机制或任务暂停。形式上,设视觉与语言模态分别为v_{vis}和v_{lang},匹配度的计算如式(6)所示。

$$\text{Score}_{\text{sim}} = \cos\left(v_{\text{vis}},v_{\text{lang}}\right) \tag{6}$$

当$\text{Score}_{\text{sim}} < \tau$时,系统认为输入存在潜在误导,进入低风险状态运行或等待用户进一步确认。此外,语义歧义检测器与语言输入模板约束机制也被广泛应用于任务的关键场景,以降低开放式自然语言对系统行为路径的不可控性影响。

综合来看,具身系统的安全机制已不再依赖某个模块的独立防护,而是朝着以“结构冗余感知、稳健策略建模、物理行为限制与语义指令过滤”为核心的全链条、多模态安全架构演进。其目标不仅在于提升个体模块的鲁棒性,更在于通过跨模块的信息一致性校验、决策路径可追踪机制与人类可干预接口,构建一

个具备协同能力、自适应调整能力与社会解释能力的综合安全系统。该系统架构的设计正逐步从“被动防御”向“主动检测、动态调整、可质询反馈”的范式转变,是具身智能走向可信部署的关键一环。

3 现实部署中的安全挑战

具身智能系统在真实场景部署中面临着复杂的安全挑战,这些挑战的特殊性源于现实世界的非理想性、高维动态环境,以及系统处理多源输入、人机协同和长期任务时的结构脆弱性与语义误差积累。

首先,现实环境的不确定性持续威胁感知鲁棒性。开放场景中普遍存在的照明变化、遮挡、物体摆放差异与传感器漂移等扰动,虽不会立即使模型崩溃,却能引发感知分布的细微漂移,进而在策略生成中诱发非线性响应,导致系统行为在边界区域出现“突变式失败”。实验室的受控环境难以暴露此类感知脆弱性。

其次,人机交互中的“语义错位”是极具隐蔽性的挑战。用户无意的模糊表达或语境不清的提示可能诱导系统偏离意图;恶意攻击者更可利用自然语言的灵活性来操纵策略。这类风险难以通过硬编码规则进行防范,更依赖系统自身的上下文推理能力与行为可解释性。

再次,在复合型多模态攻击下,具身系统暴露出跨模块协同不足的问题。视觉扰动、语言欺骗与触觉模糊等攻击模态常协同作用于感知链条的多个节点,形成“攻击路径联动效应”[13]。这要求系统具备跨模态异常检测与多通道语义一致性分析能力,当前架构在此方面普遍滞后。

此外,系统异构性是部署的结构性难题。同一模型在不同硬件平台部署时,因执行频率、控制接口、传感分辨率等差异,可能出现策略失稳或行为偏离(“策略—平台失配”)。在多平台协同时此类问题会更突出,其根源在于缺乏平台不可知的策略迁移范式和部署前一致性验证机制。

最后,文化嵌入与行为边界适配问题尤为关键。面向多文化用户群体时,系统行为规范需兼容多样价值系统,如社交距离规范、触碰接受度、语气语义理解等文化差异巨大。系统若无法动态适配这些变量,易触发社会接受障碍乃至法律纠纷。

综上,真实部署的安全挑战具有“动态生成性”与“语境耦合性”。安全机制不仅需要技术鲁棒性与容错性,更应在认知层面嵌入“社会—人类—任务”三元交互模型,其目标应超越攻击抵御与异常检测,涵盖行为解释、用户信任建构与长期稳定性的系统性保障。

4 未来研究方向

随着具身智能系统从封闭实验室逐步进入开放的社会应用场景,其安全研究也正面临深刻转型,从原本以模型鲁棒性与输入防护为主的研究重心,逐渐过渡到涵盖认知建模、社会对齐、跨模态协调与多主体共存的系统性安全架构建设,具体如表2所示。

表2 具身智能安全研究的未来关键方向概览

研究方向	核心目标	涉及技术	面临挑战
多模态鲁棒性建模	抵御组合扰动,强化感知一致性	模态对齐、跨模态注意力	模态协同机制设计困难
策略可解释与因果分析	增强用户信任与系统调试效率	因果推理、图模型、行为追踪	决策链条建模复杂
社会价值嵌入与对产	避免文化偏差与伦理冲突	偏好学习、对话建模、规则学习	规范抽象困难,评价主观
多主体协同与共识机制	构建安全可信的智能体协作系统	分布式系统、安全通信协议	风险传播路径控制难
安全评测标准与开源框架	推动横向对比与纵向复现	仿真平台、指标体系、基准数据集	社区协作与资源协调成本高

面向未来,安全机制应当更多体现“多模态融合语境中的韧性建构”。具身系统将同时面对来自图像、语言、触觉等通道的信息扰动,因此构建能够在模态间动态平衡、权重自适应的感知系统尤为重要。针对不同模态间的信息一致性问题,可引入多模态对齐损失函数,用以衡量各通道语义特征的一致程度。设视觉、语言与触觉模态分别为 x_v、x_l、x_h,对应的编码器为f_j,则一致性损失如式(7)所示。

$$\mathcal{L}_{\text{align}} = \sum_{i \neq j} \left\| f_i\left(x_i\right) - f_j\left(x_j\right) \right\|^2 \tag{7}$$

当一致性损失超过设定阈值τ时,系统将判定输入模态间存在冲突或欺骗信号,触发自检或策略回退机制。这不仅是感知模型的设计挑战,更涉及到如何建立信息通道间的冗余验证、语义一致性检测与任务一致性反馈机制。

与此同时,解释性与可追踪性将在安全性框架中扮演更加核心的角色。用户不仅要求系统“做对”,更需理解“为什么这样做”,尤其是在关键或危险情境下。透明的决策依据、因果链条与可质询路径是建立信任的基础[14]。未来,具身系统需兼具“执行正确”、

“解释清晰”与“责任明确”。

社会价值适应性成为新焦点。在多元文化、语境和人群中,系统如何理解并尊重不同社会对“合适行为”的期待,将直接影响其在教育、公共服务等领域的接受度。这不仅是技术泛化的挑战,更是规范、伦理与法律机制的协同工程。

具身智能的安全研究需突破个体,迈向群体协同。多具身体协作(如交通、应急、物流[15])要求实现跨体风险共享、策略传递与冲突预警,以提升整体韧性。此外,国际合作、标准与开源平台对安全至关重要。当前缺乏统一的评估体系与共享基准数据,这严重制约了安全机制发展。亟需共建开放测试环境与通用安全基准任务集,提升部署一致性与可信度。

未来,安全研究应以系统协同为基础,以多模态理解为支点,以人类信任为目标,构建技术—伦理—社会三位一体的综合安全生态。

5 结论

具身智能的安全保障不仅是技术问题,更关乎国家战略、产业发展与公共利益。一方面,应统筹发展与安全,推动具身智能在服务产业中的健康应用,促进智能机器人在养老、物流、教育等民生领域的可靠落地。另一方面,需坚持底线思维,优先其在保障采矿、工业、医疗等关键行业与高风险场景中的安全应用,确保技术进步不会以安全隐患为代价。此外,具身智能的安全治理离不开产学研协同机制的支撑。高校、科研机构与企业应加强合作,联合攻克具身智能的安全性、可解释性与伦理合规等核心问题,推动标准体系、评测机制与治理框架的同步构建,形成从基础理论到应用落地的闭环创新体系。

随着具身智能规模化部署的加速,构建一个“安全可信、可控可用”的智能生态,已成为技术发展与社会治理共同面对的重要课题。唯有技术安全协同,方能使具身智能从“工具”进化为“伙伴”,成为赋能产业、保障民生、增进福祉的关键基础设施。

参考文献:

[1] CANGELOSI A, BONGARD J, FISCHER M H, et al. Embodied intelligence[M]//WITOLD PEDRYCZ J K. Springer Handbook of Computational Intelligence. Berlin: Springer, 2015: 697-714.

[2] ENNALS R. Wendell wallach and colin allen: moral machines: teaching robots right from wrong[J]. AI & society, 2009, 24(2): 207-208.

[3] MITTERMAIER M, RAZA M M, KVEDAR J C. Bias in AI-based models for medical applications: challenges and mitigation strategies [J]. npj Digital Medicine, 2023, 6(1): 113.

[4] MÖKANDER J, SCHUETT J, KIRK H R, et al. Auditing large language models: a three-layered approach[J]. AI and Ethics, 2024, 4(4): 1085-1115.

[5] KIM S, LEE G G. Adversarial DPO: harnessing harmful data for reducing toxicity with minimal impact on coherence and evasiveness in dialogue agents [EB/OL]. [2025-01-03]. https://arxiv. org/abs/2405.12900.

[6] SHI D, SHEN T H, HUANG Y F, et al. Large language model safety: a holistic survey [EB/OL]. [2025-01-03]. https://arxiv. org/abs/2412.17686.

[7] LIU H P, GUO D, CANGELOSI A. Embodied intelligence: a synergy of morphology, action, perception and learning[J]. ACM Computing Surveys, 2025, 57(7): 1-36.

[8] LIU Y, CHEN W X, BAI Y J, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI[EB/OL]. [2025-01-03]. https://arxiv.org/abs/2407.06886.

[9] LAGERSTEDT E, THILL S. Multiple roles of multimodality among interacting agents[J]. ACM Transactions on Human-Robot Interaction, 2023, 12(2): 1-13.

[10] KAPELYUKH I, VOSYLIUS V, JOHNS E. DALL-E-bot: introducing Web-scale diffusion models to robotics[J]. IEEE Robotics and Automation Letters, 2023, 8(7): 3956-3963.

[11] YANG F Y, FENG C, CHEN Z Y, et al. Binding touch to everything: learning unified multimodal tactile representations[EB/OL]. [2025-01-25]. https://arxiv.org/abs/2401.18084.

[12] TURCHIN A, DENKENBERGER D. Classification of global catastrophic risks connected with artificial intelligence[J]. AI & society, 2020, 35(1): 147-163.

[13] LIU H P, GUO D, SUN F C, et al. Embodied tactile perception and learning[J]. Brain Science Advances, 2020, 6(2): 132-158.

[14] DALRYMPLE D, SKALSE J, BENGIO Y, et al. Towards guaranteed safe AI: a framework for ensuring robust and reliable AI systems: UCB/EECS-2024-45 [R/OL]. [2025-01-03]. https://www2. eecs. berkeley.edu/Pubs/TechRpts/2024/EECS-2024-45.pdf.

[15] GUPTA J K, EGOROV M, KOCHENDERFER M. Cooperative multi-agent control using deep reinforcement learning [C]//Autonomous Agents and Multiagent Systems. Cham: Springer, 2017: 66-83.

作者简介:

贾雅清,毕业于内蒙古大学,工程师,硕士,主要从事网络安全、人工智能安全研究工作;徐雷,毕业于北京理工大学,正高级工程师,博士,主要从事人工智能安全、网信安全研究工作;陶冶,毕业于英国伦敦大学,高级工程师,硕士,主要从事人工智能安全、数据安全、网信安全研究工作;王芃,毕业于清华大学,高级工程师,博士,主要负责组织公司人工智能和具身智能研发、科创生态合作等工作,长期从事物联网标准和技术研发工作。