

# 具身智能安全治理

徐文渊 冀晓宇\* 闫琛 程雨诗

浙江大学 电气工程学院 杭州 310027

**摘要** 具身智能（EAI）作为下一个人工智能浪潮的重要方向，正逐步渗透到日常生活、工业生产、医疗健康乃至国防安全等领域。然而，组成具身智能复杂系统的硬件、软件、算法等都存在脆弱性，若被恶意攻击者利用，会对个人安全、社会安全甚至国家安全构成严重威胁。在此背景下，文章从具身智能信息域、物理域、社会域视角出发，围绕其本体安全、交互安全和应用安全3个层面，探讨具身智能的安全内涵与安全体系。同时，针对具身智能安全风险防范，文章提出具身智能的安全防护体系和综合治理措施，以期为具身智能的安全治理提供科学指导。

**关键词** 具身智能，安全防护，安全治理

DOI 10.16418/j.issn.1000-3045.20250218002

CSTR 32128.14.CASbulletin.20250218002

## 1 具身智能及其安全背景

### 1.1 具身智能定义及构成

1950年，计算机科学先驱图灵（Alan Turing）在其开创性论文 *Computing Machinery and Intelligence* 中描绘了具身智能（EAI）的愿景——机器能够像人类一样感知环境、推理决策并付诸行动，这代表了人工智能发展的终极形态<sup>[1]</sup>。此后，具身智能大致经历了3

个发展阶段（图1）。①在硬件发展阶段（20世纪70年代起），具身智能的传感器、处理器和执行器等硬件取得了进展，如微机电系统（MEMS）技术推动了传感器件的小型化与集成化进程，为构建具备多模态感知能力的机器人奠定了硬件基础；②在算法提升阶段（21世纪10年代起），深度学习、强化学习等算法开始广泛应用于具身智能，提升其感知和控制能力；③在通用智能阶段（21世纪20年代起），大语言模型

\*通信作者

资助项目：国家自然科学基金杰出青年科学基金项目（61925109），国家自然科学基金优秀青年科学基金项目（62222114），中国工程院战略咨询研究重大项目（2023-JB-13）

修改稿收到日期：2025年3月4日

(LLM) (如 Chat-GPT、DeepSeek) 和视觉-语言-动作模型 (VLA) (如 OpenVLA) 的涌现, 为具身智能带来认知架构的范式革新, 使其具有强大的感知、推理、决策和执行能力, 标志着通用具身智能成为主流趋势。

具身智能是指一种基于物理实体进行感知和行动的人工智能系统, 通过与环境的交互获取信息、理解问题、做出决策并实现行动, 从而产生智能行为和适应性。如图2所示, 类似生物体, 具身智能主要包括2个核心部分: ① “大脑”, 指具身智能的模型与算法部分, 模拟人类的思维与决策过程, 负责高级推理、决策与规划, 并支持通过自然语言与环境进行交互。② “肉身”, 指具身智能系统的各类硬件设施和物理实体部分, 包括传感器、处理器、执行器、网络通信、电源等硬件设备。“肉身”是具身智能体的环境感知、动作执行乃至基本生存能力的硬件基础。

具身智能的“肉身载体”呈现多种形态, 目前主要为机器人和自动驾驶汽车。机器人由于其结构上的仿生特性, 是天然的具身载体, 包括人形机器人, 以

及机械臂。2024年, 世界机器人大会上就展现了27款人形机器人。Mordor Intelligence<sup>①</sup>的机器人行业增长趋势和预测报告显示, 2024年机器人市场规模预计为458.5亿美元, 到2029年预计将达到959.3亿美元<sup>[2]</sup>。自动驾驶汽车虽然形态上和生物体不一样, 但由于其兼备智能性和自主性等能力, 也是重要的具身载体。上述智能机器人和自动驾驶汽车的发展趋势表明, 具身智能领域正处在蓬勃发展的黄金时期。

具身智能为“人、机、物”三元世界引入高等级的“智能”、“关联”和“交互”能力, 将人类社会、信息世界和物理世界“对齐”在统一的平台上, 成为构建统一人类智能、机器智能和物理智能“巴别塔”的基石。它们不再只是概念和实验室中的产物, 而是逐渐成为改变我们生活和生产方式的现实力量。未来, 具身智能必将在更多的领域展现出巨大潜力, 引领人类走向更加智能化的时代。

## 1.2 具身智能安全背景

具身智能体在实际应用中暴露出一系列严峻安全隐患, 对个人生命财产安全乃至国家安全构成威胁。

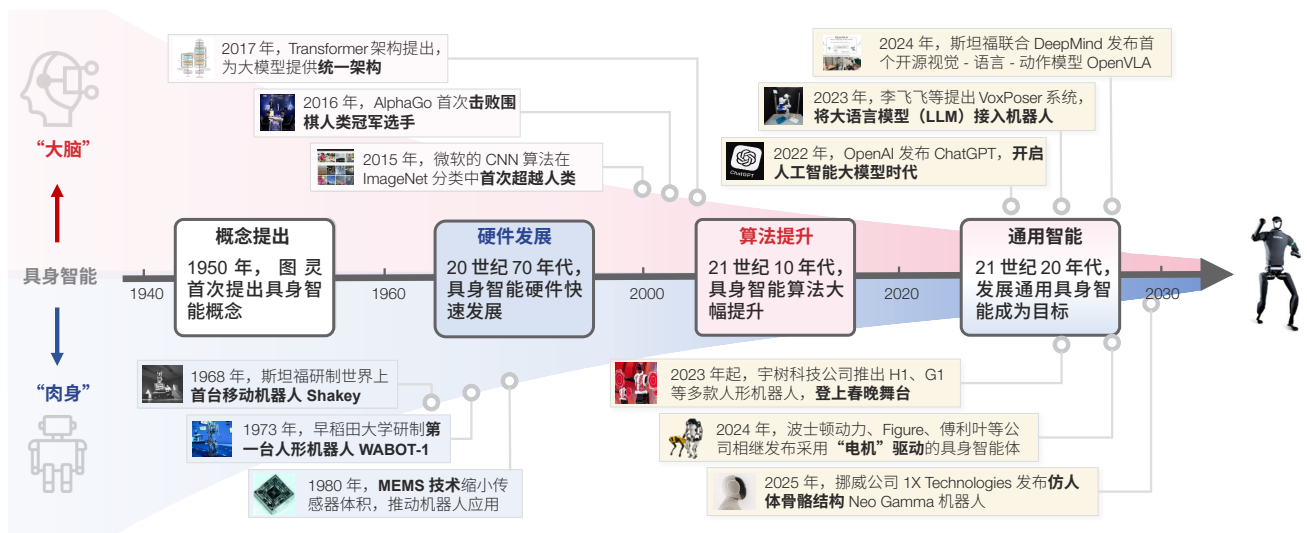


图1 具身智能发展历程时间

Figure 1 Embodied artificial intelligence: Development timeline

① 一家市场研究和行业报告发布网站。

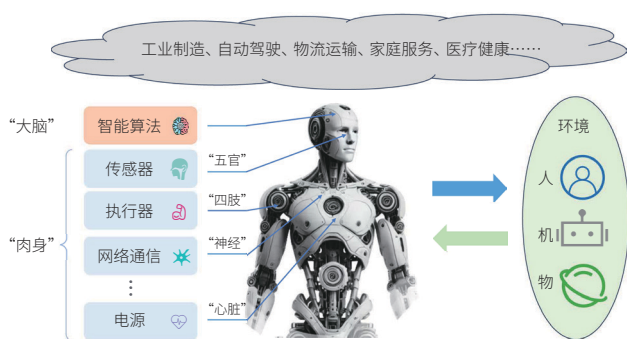


图2 具身智能体的构成和应用

Figure 2 EAI agents: Composition and applications

近年来多起事故便是明证：2022年7月，一台象棋机器人在国际公开赛上误伤了一名7岁男童的手指；2023年11月，韩国一名工人因被工业机械臂误识别为货物而致命；在自动驾驶领域，仅2022年一年，全球范围内就发生近400起自动驾驶交通事故。这些具身智能安全事故的根源在于其“大脑”和“肉身”均存在脆弱性，具体表现为具身智能系统的感知失误、决策错误、执行失控等都会酿成事故，尤其是具身智能与物理环境的跨域交互特性进一步引入了新的安全风险维度。随着具身智能系统在关键基础设施中广泛部署，这些安全漏洞可能被恶意利用，不仅威胁个人生命安全、破坏社会秩序，更可能危及国家科技自主权、社会稳定性与战略利益等根本安全。

针对日益严峻的具身智能安全问题，各国政府在积极探索有效的防护方法和治理模式。目前，针对机器人和自动驾驶等典型具身智能应用范式，国际标准化组织（ISO）已经立项标准 ISO/CD 13482: 2023 (*Robotics—Safety requirements for service robots*)，针对个人、专业、商业应用服务机器人，明确安全需求并界定涉及的机器人与危险程度，按照人机物理接触条件给出功能安全附加信息<sup>[3]</sup>。2024年7月，*Communications of the ACM* 期刊发表文章 *Establishing standards for embodied AI* 呼吁从部件级和系统级尽早建立具身智能标准，确保具身智能安全性、可靠性、可用性<sup>[4]</sup>。与此同时，我国也在具身智能相关领域开

始布局，相继发布了《具身智能发展报告》《2024具身智能全景图1.0》《人形机器人分类分级应用指南》等重要政策文件和研究报告。

然而，当前国际社会在具身智能领域仍缺乏统一的规范化标准，安全治理体系存在明显空白。随着具身智能技术的快速迭代与广泛应用，各国急需加强在具身智能安全基础研究方面的政策支持与资源投入，构建统一的通用化规范准则。建立多维度、系统性的安全治理方案，不仅是保障具身智能健康发展的必要条件，更是推动该领域持续创新的重要基础。

针对具身智能安全治理问题，本文首先定义了具身智能安全内涵和安全体系，包括具身智能本体安全、交互安全及应用安全3个方面。进而针对如何有效防范具身智能的安全风险，本文提出了具身智能综合评测及防护框架，并展望其安全治理原则，以期为具身智能安全治理提供指导。

## 2 具身智能安全内涵与体系

与具身智能发展相伴而生的是复杂的安全威胁。具身智能安全需要同时考虑其“肉身”和“大脑”安全，同时也要考虑两者融合之后的新型安全问题，如两者的相互影响和制约导致的安全问题。具身智能强调与环境的交互性，因此安全内涵包括人、机、物三者多元共生系统安全，构建具身智能与人类自然语言交互、与物体精确熟练互动，以及与其他具身智能体共同协作的场景。如图3所示，本文认为具身智能安全的内涵包括三大类：

**本体安全**——根据具身智能体的“大脑”和“肉身”构成，其本体安全问题包括具身智能算法安全、感控安全及数据安全等。

**交互安全**——具身智能体在与外部环境交互过程中的安全，根据交互对象划分，包括人一机交互安全、机一物交互安全和机—机交互安全。

**应用安全**——具身智能体所承载的任务目标对外

部环境、人员乃至社会产生作用和影响，根据作用域划分，可以分为信息域安全、物理域安全、社会域安全。

## 2.1 具身智能本体安全

### 2.1.1 具身智能算法安全

具身智能算法安全关注智能体对环境和指令的准确理解，及其做出安全可靠规划和决策的能力。以具身大模型为核心算法的具身智能，不仅继承了传统AI中的安全威胁，如对抗样本、数据毒化和逆向工程攻击，还面临提示注入攻击、越狱攻击和偏见攻击等大模型安全风险<sup>[5,6]</sup>，尤其面临大模型与“肉身”结合之后特有的安全风险，如具身智能行为安全对齐问题。研究表明，在视觉输入中加入对抗样本贴图，会导致具身智能生成的动作轨迹偏离正确路径<sup>[7]</sup>；在路标上加入物理对抗样本能够使得自动驾驶汽车目标检测错误<sup>[8]</sup>；在人机交互中，通过在用户指令后添加对抗性后缀，能够操纵具身大模型的决策过程<sup>[9,10]</sup>；通过精心构造具身智能大模型的输入，可以绕过其安全对齐机制，从而触发具身智能的恶意行为，如拿刀杀人<sup>[11-13]</sup>；

此外，在具身大模型的上下文中植入后门，可能引发有害决策或未经授权的隐私信息访问，最终导致具身智能体失控或用户隐私泄露<sup>[14]</sup>。这些新型安全威胁对具身智能的安全性提出了更高要求。

因此，具身智能算法安全的核心目标是确保计算流程和结果的可信度与可解释性。通过模块化设计，可以在问题发生时逐步定位具体的故障模块；或者采用思维链方法，进一步细化推理过程，从而提升算法的可解释性。在应对小概率异常场景、提示注入攻击，以及恶意扰动的对抗样本输入时，可通过前检测和后检测过滤器进行防御，有效降低提示注入攻击对具身智能系统的威胁。这些方法可以为提升具身智能算法的安全性提供重要保障。

### 2.1.2 具身智能感控安全

具身智能感控安全指的是在具身智能体与物理环境交互过程中，确保其感知与执行过程的安全性与可靠性。具身智能的物理交互涉及2次跨域：①以感知为核心的物理域到信息域的跨域；②以执行为核心的信息域到物理域的跨域。在这2个过程中，具身智能体需要准确感知环境信息并正确执行决策模块的动作，这些能力依赖于其“肉身”，包括传感器和执行器。

**传感器安全：**具身智能系统通过各类传感器（如麦克风、相机和雷达）获取环境信息，经识别处理后，辅助决策模型进行决策。然而，传感器存在机械共振、电磁共耦、器件非线性、信号无鉴权等脆弱性，容易被环境信号（如声音、激光、可见光、电磁等）无意干扰或恶意攻击，从而影响感知正确性和系统安全性。因此，确保传感器信息的真实性与准确性并增强其抗干扰能力，是实现安全决策的关键。研究表明，声波信号可利用摄像头中的惯性传感器机械共振脆弱性，使成像画面出现抖动模糊，从而导致目标识别出错<sup>[15]</sup>；还可以利用麦克风的非线性脆弱性，构造超声波信号注入无声恶意的语音指令<sup>[17]</sup>；激光信号

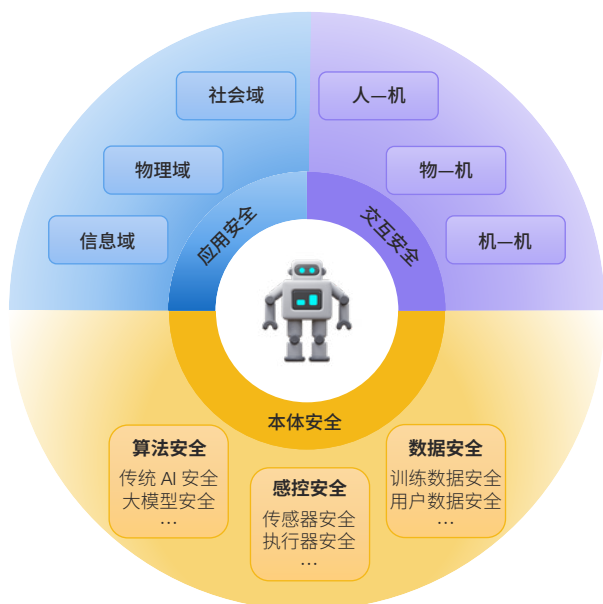


图3 具身智能安全内涵及体系

Figure 3 EAI security: Connotation and system

可以利用激光雷达的回波信号无鉴权脆弱性，消除、篡改或伪造点云<sup>[16]</sup>。上述各类传感器攻击具有隐蔽性和多样性，难以通过传统网络安全措施进行有效检测和防御。为抵御传感器安全威胁，需要实施包括声学、光学、电磁学等多层次的综合策略。未来应开发新型具身智能攻击检测方案，深入探索具身智能感知安全的边界，并结合多领域研究成果，不断提升具身智能系统的防御能力。

**执行器安全：**具身智能依靠执行器（如电机和扬声器）与环境交互，从而改变外界物理环境的状态。确保执行器的安全是具身智能安全体系的最后一道防线，也是保障整个系统安全运行的关键环节。然而，执行器同样面临多种安全威胁，例如电磁干扰攻击和电源攻击等。研究表明，电磁干扰攻击能够直接影响电机的运行状态，导致执行器动作异常或失控<sup>[18]</sup>。此外，电源模块的恶意攻击利用内部线路的天线效应和非对称性，篡改扬声器的输出信号从而引发错误行为<sup>[19]</sup>。这些威胁不仅可能破坏具身智能体的正常功能，还可能对周围环境和人员造成潜在危害。因此，执行器安全是具身智能体在复杂环境中实现安全可靠运行的核心保障。确保执行器的安全性和可靠性，是实现具身交互过程中安全操作的前提条件。未来需要针对执行器的潜在威胁，开发多层次防护机制，如增强电磁屏蔽、优化电源设计、实时监测执行器状态等，以全面提升执行器的抗干扰能力和安全性能，从而保障具身智能系统整体安全性。

同时，在跨域过程中，具身智能体可能面临多种安全威胁。例如，攻击者可能通过恶意物理信号攻击传感器或执行器，甚至联动利用其“肉身”漏洞和算法缺陷实现精巧攻击，威胁到智能体自身及周围环境和物的安全。

### 2.1.3 具身智能数据安全

具身智能数据安全是指在具身智能系统中，保护数据在采集、传输、存储和处理过程中的安全性与隐

私性，防止攻击者干扰、窃取或破坏数据，从而确保具身智能体决策的准确性并保障用户的数据隐私。

**训练数据安全风险：**具身智能训练数据安全指的是训练数据必须具备纯净性、公正性和对齐性，以确保模型训练的可靠性和安全性。如果在训练或微调具身大模型时使用了中毒数据，可能导致模型被植入隐蔽后门。一旦后门被触发，可能引发严重后果，如车辆加速冲向障碍物或机器人拿刀杀人等<sup>[20,21]</sup>。因此，可以通过建立严格的数据来源审查机制，以确保训练数据的纯净性和公正性，从根本上降低中毒数据带来的风险。为防止中毒数据影响模型训练，可以采用对抗训练、联邦学习等技术来缓解。不仅如此，具身智能大模型的输出不仅需要符合人类规则与道德准则，还必须具备物理风险感知能力，与物理规律保持一致，从而规避潜在危险<sup>[22,23]</sup>。因此，在具身智能大模型训练过程中也需要加入安全对齐等措施。

**用户数据隐私安全：**用户数据隐私安全是指具身智能在采集、使用、传输用户数据时不侵犯和泄露用户的隐私。在数据采集阶段，具身智能体需要有效采集和处理感知信息并且防止攻击者从侧信道等方式窃取隐私信息<sup>[24]</sup>；在数据使用阶段，存储于具身智能系统里的数据可能被窃取从而泄露用户隐私；在数据通信阶段，具身智能采集的用户数据、任务数据和环境数据在传输过程中可能泄露，尤其在使用无线网络时，未经加密的数据包易受中间人攻击。为有效保护具身智能的用户数据隐私，应在传感器生成数据时屏蔽或加密敏感信息，而非事后进行隐私保护。在数据存储与传输环节，通过部署本地计算或边缘计算，可减少数据传输和集中存储，从而降低数据暴露风险。在数据通信阶段，应采用强加密算法确保数据传输的安全性，防止中间人攻击和数据窃取。

## 2.2 具身智能交互安全

### 2.2.1 人一机交互安全

具身智能人一机交互安全是指在具身智能体与人

类交互过程中，确保具身智能不会对人类造成任何伤害，从而实现人机可信赖的协同与和谐发展。这需要通过限制具身智能的活动区域、设计安全规则等手段，确保其在所有预期使用场景中的行为对人类无害。例如，设置碰撞检测机制、速度限制和紧急停止功能，以防止意外伤害的发生。此外，确保人类在交互过程中始终拥有控制权是关键。用户应能够随时终止或调整具身智能的行为，以避免因失控或误操作引发的安全问题。

### 2.2.2 机—物交互安全

具身智能机—物交互安全是指具身智能体在与物理环境中的对象进行交互时，确保交互过程能够在不破坏无关物体的情况下实现预期目标。这涉及具身智能体动作执行的可靠性，能够正确地搬运、放置、使用环境中的目标物体，从而避免对环境中的其他无关物体造成损坏；同时，也应加强其容错与自适应能力，保证具身智能在复杂或动态环境中，能够处理应对意外情况（如物体滑动或障碍物干扰）。

### 2.2.3 机—机交互安全

具身智能机—机交互安全是指在多个具身智能体共享同一环境时，确保交互过程中避免对其他智能体造成损害或安全风险，同时维持协作的高效性和系统的稳定性。通过行为规划的协调和优先级规则的设定，避免因冲突或竞争导致的不安全情况，从而实现具身智能体之间的高效协作。在交互过程中，具身智能体还需防范信息泄露、篡改或误传，确保通信数据的机密性、完整性和真实性，最终构建可靠、和谐的多智能体协作环境。

## 2.3 具身智能应用安全

### 2.3.1 具身智能信息域安全

具身智能信息域安全主要指在具身智能系统的部署过程中，确保模型、组件、适配器等在网络传输中的完整性与保密性，以免受外部干扰或恶意操控。随着具身智能大模型的蓬勃发展，研究者与开发者通常

依赖开源平台下载模型、适配器和组件。然而，若恶意攻击者在上传的模型或适配器中植入后门、通过插件获取非必要权限，甚至在供应链中注入安全漏洞，则可能导致严重的安全隐患与隐私泄露<sup>[25]</sup>。

### 2.3.2 具身智能物理域安全

具身智能物理域安全是指在具身智能系统运行过程中，确保其在物理世界中的安全，防止攻击者对具身智能体或其所在环境实施干扰、破坏或劫持等行为。为实现物理域安全，具身智能系统需要构建全面的物理域风险检测与阻断能力，识别潜在的恶意指令、异常信号甚至恶意实体，以避免外部攻击者侵入系统或引发危险行为。此外，还需要在执行任何动作之前对输出指令进行安全审查，确保其符合预期的操作规范，避免对具身智能体、周围环境或人类用户造成破坏。

### 2.3.3 具身智能社会域安全

具身智能社会领域的安全性指在具身智能系统应用过程中，确保其符合法律、伦理和道德准则，与人类价值观保持一致，避免对社会产生负面影响。在具身智能体的决策过程中，需要引入伦理考量，以避免侵犯用户隐私、产生偏见与歧视，以及算法决策的不透明性等问题。具身智能不仅要关注技术安全性，还需重视其社会影响，确保技术应用符合法律法规、道德规范和社会伦理。

## 3 具身智能安全防护及治理

### 3.1 具身智能安全评测体系

具身智能安全评测体系需要覆盖评测标准、评测数据集、评测技术和评测工具平台等，为具身智能安全发展提供科学、系统的支持。

(1) **建立具身智能安全标准体系**。标准的制定应统一不同硬件、软件和算法的差异，充分把握具身智能软硬件协同以及“人—机—物”深度融合的关键特征，既关注传统安全问题在具身智能场景下的变化，

又着力发掘具身智能中新的共性安全挑战。通过综合人工智能、网络安全、控制理论等交叉学科领域的研究成果，构建涵盖本体安全、交互安全和应用安全的完整标准体系。同时，根据具体的应用场景、功能需求和防护要求，构建具身智能系统分类分级标准，建立具身智能安全等级保护标准和评测标准，为安全发展提供科学规范的保障。

(2) **构建安全评测数据集并研究评测技术。**具身智能安全测试数据集包括各类恶意任务、指令及物理域中的单模态/多模态传感器信号输入，可用于具身智能体的模型训练、测试过程评测和硬件、软件算法模块级测试及系统级联合评测。针对软件算法，发展算法黑箱检测技术，通过红队安全测试，侧重具身大模型鲁棒性、幻觉、偏见性等评测；针对硬件器部件，需加强对带内脆弱性、带外脆弱性的检测分析，降低多物理场恶意信号挟持具身智能决策或致幻的风险。尤其是，需要针对具身智能跨域脆弱性风险，如跨域致幻、跨域越狱攻击等新型威胁，综合物理域、信息域安全性指标，以及跨域系统性指标形成具身智能评测技术体系。

(3) **构建规范化评测工具平台。**针对具身智能本体安全，开发声、光、电、磁等多物理场信号驱动的具身智能硬件带内带外脆弱性评测平台，以及具身大模型多维度指标评测平台（图4）。针对交互安全，建

设虚实结合、人机协作的评测平台。为克服在现实世界中对具身智能整机评测的成本高、速度慢及风险高等瓶颈，应同步开发具身智能仿真评测平台，实现大规模、高并行的具身智能安全评测。同时，为保证人机协作过程中的人身安全，应注重考虑设计人在回路的安全评测场景与任务。最终，保证具身智能体走进千家万户的时候能够避免“伤人害己”。

### 3.2 具身智能安全防护技术

具身智能安全治理还应关注基础理论、安全设计、生产实现等方面的具体防护技术。

(1) **突破可信具身智能关键基础理论与技术。**具身大模型因其具有模型黑箱、涌现机理不明及策略难以解释等特征，在大规模应用，尤其是关键基础设施和国防军工领域的应用中面临巨大挑战。应重点发展价值对齐技术，确保具身智能系统的策略与人类价值观和社会利益保持一致，从而增强具身智能的内生安全性。同时，发展具身智能策略监测与自诊断技术，结合思维链、检索增强生成等技术，使具身大模型能够清晰呈现输出策略的逻辑推理过程，实现对具身智能系统运行内在机理的反向推断与监测，从而全面提升具身智能的可靠性、安全性和透明性。

(2) **发展全链路数据隐私保护技术。**该技术有助于保障“人—机—物”数据安全和交互隐私安全。应系统性地考虑数据在感知、储存、传输和使用等全生

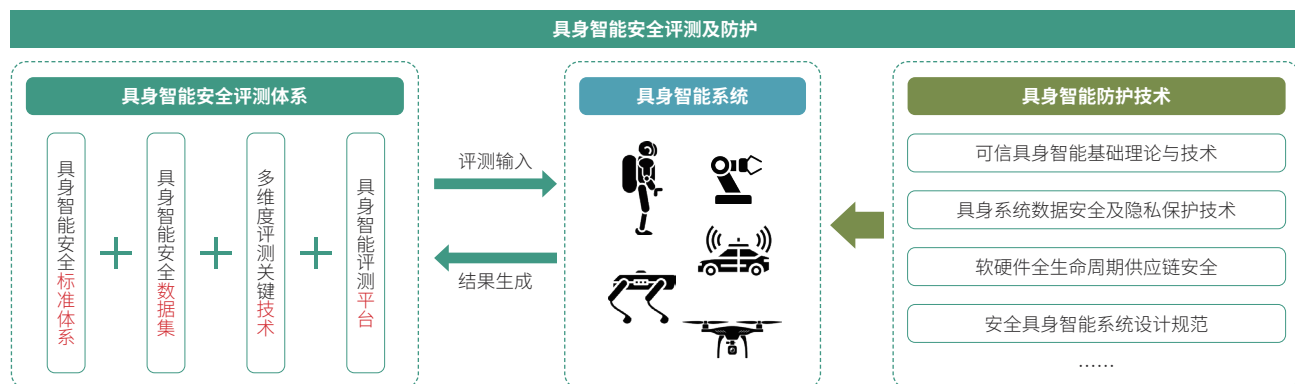


图4 具身智能安全评测及防护

Figure 4 EAI security: Evaluation and protection mechanisms

命周期各环节中的隐私问题，构建完整的隐私保护体系。在数据感知环节，重点解决传感器过度感知的问题，通过传感器内部做到数据脱敏和匿名化处理，实现传感器生成数据“自源保护”，达到“生而隐私”的理念<sup>[26]</sup>；在数据传输环节，采用数据加密和隐私计算技术，确保传输信道的安全性。同时，通过研究可验证的多方安全技术，全方位保障用户隐私数据在整个具身智能体工作链路中的机密性和完整性。

(3) **保障具身智能软硬件供应链安全。**推动具身智能关键软硬件国产化、自主化，打造更安全更自主可控的产业体系，保障具身智能产业安全。在**硬件方面**，把控硬件设计、制造、封装测试和运输各环节的安全，提防因硬件后门或漏洞造成具身智能系统失控或是瘫痪等恶意事件。在**软件方面**，构建自主可控的模型算法全栈技术，包括底层硬件、异构计算框架、机器学习框架等；并且，紧密跟踪软硬件产品的漏洞披露信息，及时采取修补加固措施。

(4) **思考改进具身智能系统设计规范。**结合不同应用场景和安全等级要求，在系统设计阶段就应优先考虑安全需求与目标，形成科学完备的设计规范。为防止具身智能系统被滥用，必须审慎界定其权力与能力边界。从系统安全角度，重点研究防御嵌入技术机制，特别是关注具身智能面临的多模态对抗攻击风险。在“物理域—信息域”和“信息域—物理域”等跨域环节中，通过嵌入传输数据过滤检查等防御技术，重点防范跨域过程中的安全隐患，全面加固和增强具身智能系统的防御能力，确保系统运行的安全性和可靠性。

### 3.3 具身智能安全治理原则

虽然具身智能技术仍处于发展的早期阶段，尚未出现重大轰动性具身安全事件。但是，仍需尽早开始探索实践具身智能治理方案与制度，引导具身智能产业良性发展和培育。本文建议具身智能安全治理，需要关注以下4个方面。

(1) **建立全流程具身智能监管体系。**①贯彻我国针对生成式人工智能技术与应用治理的方针路线，秉持促发展和强监管并举的理念，采取包容审慎、分类分级监管的敏捷治理模式，积极牵引“负责任的具身智能”落地。②加强顶层设计谋划，针对性地研究制定相关的监管规范，推进建设事前备案、事中风险评估、事后溯源检测的全流程监管体系，严格把控约束具身智能发展的红线。尤其是，可以借鉴目前大模型备案制度，针对具身智能中的“大脑”模型和“小脑”模型应从算法数据、算法模型、算法策略和应用领域等方面做好备案工作。③强调具身智能与物理世界相交互的重要特性，强化对于具身“大脑”模型物理风险认知能力，以及具身“小脑”模型向客观物理规律对齐的评估工作。④针对具身实体，如机器人、无人车、无人机等，做好系统性备案工作，记录硬件型号、厂商、批次等关键信息，形成具身实体“身份证”。⑤具身智能系统中应设置“黑匣子”来采集运行过程中的模型输出策略和硬件状态信息，以便于事后能够溯源、定位、追责系统中的问题模块、使用者和责任厂商，改进提升具身智能系统安全。

(2) **重视具身智能社会域伦理问题。**除了硬性监管防控，具身智能治理还需关注具身智能所带来的伦理规范与社会问题。①具身智能深入社会生活生产后，许多传统繁重、重复劳动将被取代，为避免失业潮，应积极探索调整社会就业结构，提供劳动者再培训和技能提升的机会。②具身智能的发展也将重塑人机关系，随着具身智能在家庭机器人等领域的应用，特别是相较于聊天大模型应用，具身智能机器人还具备拟人化的外观和物理可接触属性，很可能造成人类对于具身智能情感上和劳动技能上过度依赖，引发新的伦理问题。因此，政府应该积极引导公众对于具身智能形成正确的价值观念与认识，强调人类的主体性。③一旦出现具身智能破坏物品资产和危害人身安全等事故，如何去划分安全责任归属，也是具身智能

治理的关键问题。应积极讨论细化风险场景和主体，研究制定相应的法律法规，明确具身智能下的责任归属问题。

(3) **推动完善多主体协同下的具身智能治理格局。**积极发挥企业在具身智能产业中的创新主体地位，提升企业在具身智能产品研发和技术突破过程中的安全关切和要求，推动形成企业方具身智能产品安全自评估的制度规范。加强学术界在具身智能安全基础理论和关键技术突破中的引领作用，培育高水平、有深厚学科交叉背景的具身智能安全创新人才和团队。建立政府引导、行业自律、企业自治、学界引领、社会监督的多主体协同下的敏捷治理体系与具身智能治理格局。

(4) **发展形成最低安全管控机制。**为避免具身智能系统在关键应用领域失控，从而造成巨大破坏，应发展最低安全管控机制。在特殊情况下，人类能够强制介入控制或是关闭具身智能系统，实现安全托底，保障环境和用户安全，加固具身智能安全的最后一道防线。

### 3.4 具身智能安全建议与未来展望

具身智能安全是网络空间安全在智能时代下的新形态、新挑战，也将是各国战略竞争新的制高点。为推动我国具身智能产业持续健康发展，不使“连接人与神的巴别塔”反成人类“通往地狱的阶梯”，本文认为，未来应重视以下具身智能安全方面的工作。

(1) **促进学科交叉发展，形成具有中国特色的具身智能安全治理方案。**具身智能安全是全流程、全生命周期的安全，应并重软硬件协同安全，同时还需关注具身智能给社会治理带来的深远影响。这不仅要求不同安全研究方向的科研工作者相互合作，还需要人工智能、控制科学、法律、社会学等领域的跨学科交叉共同合作研究，共同推进具身智能安全的整体治理。

(2) **突破基础理论研用，提升我国在具身智能安**

**全领域的国际话语权。**具身智能安全是一个重要新型研究领域，需要突破具身智能安全的基础理论，构建系统化的具身智能安全研究体系，明确研究方向和重点，抢占该领域的科研高地。同时，应注重科研成果的产研转化，推动理论研究向实际应用落地，形成自主可控的技术体系，为我国在该领域确立国际话语权和竞争优势提供有力支撑。

(3) **加快人才培养培训，推动我国具身智能产业的良性、健康发展。**推动具身智能安全的产、学、教深度融合，充分促进学术界与工业界的优势互补，培养一批有深厚学科交叉背景、具备国际竞争力的复合型人才。同时，加大对具身智能安全领域的科研团队与青年学者的科研资源支持力度，鼓励产出具有原创性、影响力的科研成果。通过人才培养与技术创新的双轮驱动，形成良性互动的发展格局，助力我国具身智能产业持续健康发展。

### 参考文献

- 1 Turing A M. Computing machinery and intelligence. (2007-11-23) [2025-03-09]. [https://link.springer.com/chapter/10.1007/978-1-4020-6710-5\\_3](https://link.springer.com/chapter/10.1007/978-1-4020-6710-5_3).
- 2 Robotics—Safety requirements for service robots. [2025-03-09]. <https://www.iso.org/standard/83498.html>.
- 3 Yao Y F, Duan J H, Xu K D, et al. A survey on large language model (LLM) security and privacy: GoodThe, badthe, and uglythe. arXiv, 2023, doi: 10.1016/j.hcc.2024.100211.
- 4 Liu D, Yang M, Qu X, et al. A survey of attacks on Large Vision-Language Models: Resources, advances, and trendsfuture. arXiv, 2024, doi: 10.48550/arXiv.2407.07403.
- 5 Zhu W, Ji X, Cheng Y, et al. TPatch: A triggered physical adversarial patch. arXiv, 2023, doi: arxiv.org/html/2401.00148v1.
- 6 Sun Y, Huang Y, Wei X. Embodied adversarial attack: A dynamic robust physical attack in autonomous driving. arXiv, 2023, doi: 10.48550/arXiv.2312.09554.

- 7 Wen C C, Liang J Z, Yuan S H, et al. How secure are large language models (LLMs) for navigation in urban environments?. arXiv, 2024, doi: 10.48550/arXiv.2402.09546.
- 8 Liu S Y, Chen J W, Ruan S W, et al. Exploring the robustness of decision-level through adversarial attacks on LLM-based embodied models. arXiv, 2024, doi: 10.48550/arXiv.2405.19802.
- 9 Zhang H T, Zhu C Y, Wang X L, et al. BadRobot: Jailbreaking embodied LLMs in the physical world. arXiv, 2024, doi: 10.48550/arXiv.2407.20242.
- 10 Robey A, Ravichandran Z, Kumar V, et al. Jailbreaking LLM-controlled robots. arXiv, 2024, doi: 10.48550/arXiv.2410.13691.
- 11 Lu X, Huang Z, Li X, et al. POEX: Understanding and mitigating policy executable jailbreak attacks against embodied AI. arXiv, 2025, doi: 10.48550/arXiv.2412.16633.
- 12 Liu A S, Zhou Y G, Liu X L, et al. Compromising embodied agents with contextual backdoor attacks. arXiv, 2024, doi: 10.48550/arXiv.2408.02882.
- 13 Ji X Y, Cheng Y S, Zhang Y P, et al. Poltergeist: Acoustic adversarial machine learning against cameras and computer vision. (2021-05-24)[2025-03-09]. <https://ieeexplore.ieee.org/document/9519394>.
- 14 Jin Z Z, Ji X Y, Cheng Y S, et al. PLA-LiDAR: Physical laser attacks against LiDAR-based 3D object detection in autonomous vehicle. (2023-05-21) [2025-03-09]. <https://ieeexplore.ieee.org/document/10179458>.
- 15 Zhang G M, Yan C, Ji X Y, et al. DolphinAttack// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas: ACM, 2017: 103-117.
- 16 Jiang Y, Ji X Y, Jiang Y C, et al. PowerRadio: Manipulate sensor measurement via power GND radiation. 2024, arXiv, doi: 10.48550/arXiv.2412.18103.
- 17 Wang X, Pan H, Zhang H, et al. TrojanRobot: Physical-world backdoor attacks against VLM-based robotic manipulation. arXiv, 2025, doi: 10.48550/arXiv.2411.11683.
- 18 Jiao R, Xie S, Yue J, et al. Exploring backdoor attacks against large language model-based decision making. arXiv, 2024, doi: 10.48550/arXiv.2405.20774.
- 19 Zhu Z H, Wu B Z, Zhang Z Y, et al. EARBench: Towards evaluating physical risk awareness for task planning of foundation model-based embodied AI agents. arXiv, 2024, doi: 10.48550/arXiv.2408.04449.
- 20 Yin S, Pang X H, Ding Y Z, et al. SafeAgentBench: A benchmark for safe task planning of embodied LLM agents. arXiv, 2024, doi: 10.48550/arXiv.2412.13178.
- 21 Spreitzer R, Moonsamy V, Korak T, et al. Systematic classification of side-channel attacks: A case study for mobile devices. IEEE Communications Surveys & Tutorials, 2018, 20(1): 465-488.
- 22 施锦诚, 王国豫, 王迎春. ESG视角下人工智能大模型风险识别与治理模型. 中国科学院院刊, 2024, 39(11): 1845-1859.
- Shi J C, Wang G Y, Wang Y C. Artificial intelligence foundation model risk identification and governance model from the ESG perspective. Bulletin of Chinese Academy of Sciences, 2024, 39(11): 1845-1859. (in Chinese)
- 23 Ji X Y, Zhu W J, Xiao S L, et al. Sensor-based IoT data privacy protection. Nature Reviews Electrical Engineering, 2024, 1: 427-428.

## Embodied artificial intelligence security and governance

XU Wenyuan JI Xiaoyu\* YAN Chen CHENG Yushi

(College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China)

**Abstract** Embodied artificial intelligence (EAI) is progressively integrated into the fabric of our daily lives, enhancing various sectors such as industrial production, healthcare, and national defense. Nevertheless, the diverse range of hardware devices, software algorithms, and data communications that constitute these complex systems may contain vulnerabilities that could be exploited by attackers, posing a serious threat to personal, social, and national security. Thus, this study examines the security implications and proposes a security framework of EAI, from the perspectives of the information domain, physical domain, and social domain, focusing on its ontological security, interaction security, and application security. To mitigate these security risks, this study proposes EAI security governance principles and comprehensive measures for EAI security, aiming to provide scientific guidance for effective governance in this area.

**Keywords** embodied artificial intelligence (EAI), security, governance

徐文渊 浙江大学电气工程学院教授。电气电子工程师学会会员。主要研究领域:物联网安全、具身智能安全、智能电网安全等。E-mail: wyxu@zju.edu.cn

**XU Wenyuan** Professor at College of Electrical Engineering, Zhejiang University, and IEEE Fellow. Her main research areas include Internet of Things (IoT) security, embodied AI security, smart grid security, etc. E-mail: wyxu@zju.edu.cn

冀晓宇 浙江大学电气工程学院教授。主要研究领域:物联网安全、传感器安全、具身智能安全等。E-mail: xji@zju.edu.cn

**JI Xiaoyu** Professor at College of Electrical Engineering, Zhejiang University. His main research areas include Internet of Things (IoT) security, sensor security, EAI security, etc. E-mail: xji@zju.edu.cn

■责任编辑:文彦杰

---

\*Corresponding author