具身智能安全：内涵及治理

——第二十一期CCF秀湖会议报告

整理者：陆炫存　黄郑献　郑一凡　张世琦　应天棋　王　禹　李鑫宇

背景与意义

随着机器人、大模型技术与物理实体的融合，“DeepSeek+ 宇树”模式具身智能体等被视为下一波人工智能创新的浪潮。构成具身智能复杂系统的硬件、软件、算法和数据，都可能成为恶意攻击者利用的对象，从而在现实世界中引发危害，对人身安全和用户隐私构成严重威胁。在此背景下，研究学者和开发者迫切需要全面关注具身智能系统的整体安全性，并在研究和应用过程中考虑其中的技术、法律、伦理和道德等问题。

针对上述问题，CCF 组织了第二十一期秀湖会议，邀请了 20 余位学术界和产业界的专家学者围绕“具身智能安全”这一主题开展观点报告和集中研讨。根据具身智能安全研究中的关键技术及其安全的研究视角，会议组织了五个专题讨论。

本文总结梳理了本次会议学术交流和思想碰撞中产生的关于具身智能安全的观点和认识，梳理了已有的研究实践，并形成了对未来技术发展的思考与展望，初步达成了关于具身智能安全的内涵、防护治理等方面的共识，并发出了相应倡议。

专题一：具身智能及其安全总体展望

具身智能是指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、做出决策并付诸行动，从而产生智能行为和适应性。随着多模态大模型技术的发展，具身实体如机器人、机械臂、自动驾驶汽车等可以利用大模型，获得强大的泛化、理解和推理能力，以及物理世界中的各类复杂操作能力。具身智能概念受到工业界和学术界的广泛关注与研究，特别是拟人化的人形机器人成为目前火热的技术前沿。

然而，由于具身智能系统同时存在于物理域和

信息域，涉及复杂的软硬件控制，相较于一般的大模型安全问题，具身智能还面临更多来自物理域的攻击威胁，例如硬件漏洞、恶意信号干扰等，安全挑战尤为严峻。此外，具身智能具有与物理现实交互的特性，一旦系统失控，可能对环境、设备及人员造成严重危害，甚至威胁社会和国家安全。因此，具身智能的安全问题已成为学术界和工业界亟待研究的重要课题。与会专家学者从多个角度分享了具身智能技术发展及其安全性的研究成果、实践经验、观点以及未来展望。

西湖大学教授金耀初的报告聚焦于集群机器人系统中的安全与隐私保护这一核心科学问题。当前研究状况表明，尽管部分传统隐私保护技术（如差分隐私、联邦学习）已被尝试应用于该领域，但针对集群机器人特有属性（如分布式协作、动态拓扑、资源受限）的研究仍处于初步探索阶段。一个关键的技术挑战在于集群系统固有的“黑箱”特性，即可解释性严重不足，这直接阻碍了对其安全性和可靠性的评估与保障。因此，提升集群机器人决策与行为的可解释性成为了重要的研究前沿。此外，如何在保障高等级安全与隐私需求的同时，维持集群系统执行任务所需的实时性与效率，以及如何在人机共融场景下建立可信赖的人－集群交互机制，都是亟待解决的科学难题与未来重要的研究方向，涉及优化理论、控制论、可信人工智能等多个交叉领域。

西北工业大学教授王震探讨了基于大模型的分布式智能系统的感知、决策与控制一体化安全问题。他凝练出三个关键科学挑战：其一，空间智能的维度提升，即如何将大模型强大的语义理解与推理能力从二维信息空间有效扩展至复杂的三维物理环境进行感知与交互，其中涉及多模态融合、3D 场景理解与重建等前沿技术；其二，算法工程的效能与精度优化，即如何设计高效、精准且资源消耗合理的算法，使具身智能体能够在物理世界中可靠地完成复杂任务，这关乎计算效率、任务规划与泛化能力；其三，动态环境下的鲁棒控制，即如何在非结构化、不确定性环境中确保具身智能体稳定安全地执行任务，这对控制理论与强化学习等领域提出了更高要求。王震进一步强调，具身智能的安全是一个系统工程问题，必须建立贯穿从底层硬件供应链到上层应用软件乃至核心算法的全链路安全保障体系，这为未来研究指明了多学科交叉、全方位防御的技术方向。

上海交通大学教授朱浩瑾深入剖析了信息物理系统（Cyber-Physical System, CPS）融合背景下具身智能引发的新型安全威胁范式。其中的核心科学问题是：具身智能的引入如何打破传统信息域与物理域的安全边界，使网络攻击得以直接映射为物理世界的实际危害。这一跨域特性对安全研究提出了全新的挑战。朱浩瑾从本体感知安全（如传感器欺骗与对抗攻击）、智能决策安全（如模型后门、数据投毒对决策逻辑的篡改）和硬件执行安全（如执行器控制指令的劫持与篡改）三个维度，系统性地解构了具身智能面临的端到端安全风险链条。他强调在智能物联网（AIoT）场景下，大模型适配器（adapters）的安全问题成为大模型在资源受限设备上安全、有效部署的关键技术瓶颈，是当前嵌入式 AI 安全和可信 AI 领域亟待攻克的前沿研究课题。

宇树科技技术总监何银军从产业实践视角揭示了人形机器人作为具身智能前沿载体的技术演进路径与工程挑战。通过对行业现状，特别是对特斯拉 Optimus-Gen2 和波士顿动力 Atlas 电驱版等代表性平台的技术解构，何银军阐述了当前驱动技术、运动控制、感知系统以及通用人工智能集成等方面的研究热点与技术前沿。实例展示环节呈现了将通用人工智能技术落地于物理机器人时所面临的工程难题与应用前景，突显了仿真到现实（Sim-to-Real）迁移、模型部署优化、实时决策与鲁棒执行等关键科学与工程问题。何银军对产业链上、中、下游的分析，不仅为理解该领域的技术生态提供了框架，也间接指出了从核心零部件创新（上游）、系统集成与软件开发（中游）到场景应用落地（下游）各环节蕴含的技术突破点与未来研究方向，为学术界与产业界的协同创新提供了有价值的参考。

专题二：具身智能算法及控制安全

具身智能算法与控制安全是指在具身智能体与

物理环境交互过程中，如何确保决策与执行的行为安全可靠。具身智能涉及机器人、自动驾驶以及智能制造等高动态性的重要场景，面临决策正确、实时、安全和控制鲁棒等多重挑战。而随着具身智能技术的发展，具身智能系统控制的安全性和可靠性对实际应用的影响愈加显著，成为学术界和工业界关注的焦点。与会专家学者从不同角度探讨了具身智能算法与控制的安全风险，分享了各自的研究进展、实践经验及未来的研究方向。

清华大学教授陶建华的报告聚焦于 AI 大模型时代背景下的核心科学挑战与前沿技术方向。报告指出，大模型有望成为类似操作系统的基础性平台，但其广泛应用面临两大关键科学问题：可信性和可解释性。针对可信性问题，当前的研究前沿聚焦于探索知识图谱与大模型的深度融合机制，旨在利用知识图谱的结构化知识增强大模型输出的事实一致性与可靠性。对于可解释性难题，报告探讨了利用探针、对抗性攻击和可视化等技术手段，解构大模型复杂的内部运作机制，探究其强大能力来源的可能性，这代表了理解和剖析大模型黑箱的重要研究方向。此外，报告强调了大模型构建和应用中固有的安全脆弱性问题，并展望了未来研究须加强人工智能安全技术研发和伦理治理机制构建，特别是针对大模型生成内容的自动化审核与规范化技术，是保障大模型安全可控发展的关键研究议题。

湖南大学教授李树涛的报告深入探讨了具身智能领域的核心科学问题与技术进展。报告认为，实现自然高效的人机交互是具身智能成功的关键，这构成了该领域的核心科学挑战。在此背景下，多模态信息融合和人机交互过程中的安全性成为亟待突破的技术瓶颈。报告展示了当前研究前沿在相关方向上取得的进展，包括：（1）情感与意图识别，通过多模态情感分析提升交互的自然性；（2）复杂场景理解，如多人交互行为理解、开放域视觉问答和交互式视频定位，这些是提升具身智能体环境感知与交互能力的关键技术；（3）高效部署与安全，解决模型在实际应用中的效率和安全保障问题。报告特别指出了大语言模型为具身智能带来的新机遇，预示着利用大语言模型进行高级指令生成与理解、驱动更深层次的多模态学习将是未来重要的研究前沿。同时，模型压缩与端云协同架构下的具身智能系统设计与优化，也是未来实现技术落地必须解决的关键问题。

西安交通大学教授沈超的报告聚焦于人工智能原生安全与隐私风险这一新兴交叉领域。报告首先明确了 AI 系统，特别是大模型和具身智能，所引发的安全风险已超越传统网络安全范畴，对个人、社会乃至国家安全构成潜在威胁，这是一个亟待系统性研究的宏观科学问题。报告从信息安全的经典 CIA（机密性 Confidentiality，完整性 Integrity，可用性 Availability）三要素出发，深入剖析了大模型和具身智能在这三个维度上面临的独特风险与挑战，例如模型窃取、数据投毒、对抗样本攻击等，为理解 AI 系统的内在脆弱性提供了理论框架。报告分享了面向移动端 AI Agent 系统进行风险矩阵构建与量化评估的研究实践，代表了在特定应用场景下进行 AI 风险评估与管理的技术前沿探索。展望未来，报告指出了几个关键研究方向：（1）构建系统化的 AI 安全评估框架与基准；（2）拓展和验证可信 AI 在现实复杂场景中的应用；（3）提升 AI 系统的可解释性，并将其作为增强安全性和可靠性的重要途径。这些方向共同构成了确保 AI 技术安全、可靠、负责任发展的核心研究议题。

浙江大学教授王志波的报告着重探讨了端云协同计算架构下大模型部署与应用所面临的效率与安全双重挑战。报告指出，随着模型参数规模达到万亿级，如何在资源受限的智能终端上有效部署和运行大模型，成为一个关键的工程与科学问题。端云协同被认为是当前最有前景的技术路径，但这引入了新的复杂性，具体表现为模型部署策略、跨设备资源调度优化，以及端云计算任务协同机制等方面的研究挑战。在安全层面，这种分布式架构加剧了隐私泄露、内容可信度和数据遗忘等风险，对安全协议和机制设计提出了更高要求。报告介绍了面向大模型应用场景的自适应红队测试方法，作为评估和提升大模型鲁棒性的前沿技术探索。同时，报告

揭示了大模型遗忘学习机制可能引入新的隐私泄露风险，这是一个值得深入研究的潜在安全隐患。展望未来，研究重点将聚焦于：（1）研发轻量化的端侧大模型；（2）设计安全高效的端云大模型协同学习与推理框架；（3）探索端云协同环境下的综合安全保障技术。

腾讯 AI Lab 研究员吴秉哲的报告聚焦于基于可信智能体的开放域风险治理这一前沿研究议题。报告首先明确了在虚拟及物理世界广泛应用前景下智能体落地面临的严峻挑战，包括合规性风险、感知决策的可靠性以及面对恶意网络攻击时的脆弱性，这些共同构成了构建可信智能体的核心科学问题。为系统应对这些挑战，报告提出了构建可信智能体的四大关键技术要素，代表了当前该领域的重要研究方向：（1）风险智能感知，发展智能体自主识别、理解和预测潜在风险的能力；（2）可靠性增强范式，研究系统性提升智能体行为鲁棒性、稳定性和安全性的方法论与技术；（3）推理可解释性，确保智能体决策过程透明、可被人类理解、易于审查与调试；（4）全链路数据隐私保护，在智能体数据采集、处理、存储、使用的各个环节提供严格的隐私保障机制。

在集中讨论环节，与会专家就具身算法及控制安全面临的挑战和已有实践经验进行了充分探讨与交流。在讨论了现有的具身智能算法及控制框架的安全性后，与会专家们一致认为，具身智能端到端架构具有广阔的前景，且在特定的场景中已经展现出了优势，但由于缺乏机器人数据，技术还需一定时间的发展与研究。目前，大小脑协同的分层决策框架，在开放环境、智能驾驶等复杂场景中，取得了优异的效果。除此之外，专家就具身算法及控制安全达成了如下共识：

- **关注具身智能端侧模型的安全风险：**大模型在具身终端部署，往往需要通过模型压缩、剪枝以及量化等技术手段，这些模型轻量化手段给模型的安全性造成的影响尚未充分研究。因此，针对端侧模型还需严格的安全评估，并采取安全护栏等措施，如前置或后置安全评估，确保模型安全。
- **须明确具身智能的能力边界：**针对具身智能的不同任务，应在设计初期考虑安全因素，在特定任务上设定明确的功能需求与能力边界，结合本质安全的理念，确保具身智能系统在处理关键或高风险任务时具备安全约束与防护措施。
- **推动法律法规的完善：**目前，针对具身智能安全归属问题，在法律上还存在一定的空白。例如，自动驾驶车辆发生事故时，谁来担责的问题尚未明确。因此，须倡导建立相应的法律法规，明晰责任的归属，以便于具身智能良性发展。

专题三：信息物理系统融合下具身智能安全

信息物理系统通过集成先进的感知、计算、通信、控制等信息技术和自动控制技术，构建了物理空间与信息空间中人、机、物、环境、信息等要素相互映射、适时交互、高效协同的复杂系统，实现系统内资源配置和运行的按需响应、快速迭代、动态优化。具身智能是典型的信息物理系统，其安全存在复杂性和多样性，与会专家各自分享了对信息物理系统融合下具身智能安全的观点、看法和实践经验。

中国科学院计算技术研究所研究员蒋树强的报告探讨了具身智能的本质及其安全涵义。报告首先提出了一个核心科学观点：智能的产生与具身物理体验存在内在的、不可分割的联系，因此必须将两者联合研究。具身智能系统可被解构为决策规划“大脑”、运动控制“小脑”和硬件物理“本体”三个相互耦合、相互支撑的核心组成部分，如何实现这三者之间的高效、安全协同，构成了系统层面的关键科学问题。具身智能作为人工智能从信息空间迈向物理世界的关键载体，其安全性是一个贯穿硬件本体、物理环境、人机交互乃至社会伦理的广义、多维度概念。从技术角度出发，具身智能安全涉及涉身性、交互性、自主性、情境性、社会性等多个维度，并强调需要超越单一算法安全，研究多智能能力（如感知、问答、行为、推理、学习）复合下的安全性，这是一个更复杂且亟待解决的科学挑战。以具身导航为例，报告生动阐释了感知准确性（“眼不能歪”）、决策鲁棒性（“脑不能乱”）、学习无偏性（“学不能偏”）、执行可靠性（“脚不能跛”）等具体安全要素，

指明了具身智能安全既包含传统 AI 安全的共性，又具有独特的、内涵丰富的特性，是未来研究的重点领域。

西安交通大学教授苏洲的报告以移动群智感知作为研究具身智能安全的切入点，聚焦于分布式、移动化场景下的安全挑战。报告识别了三类典型的攻击模式，揭示了该场景下的核心安全科学问题：（1）通信链路脆弱性，针对通信功能的干扰攻击；（2）感知数据污染，利用智能移动组件（如恶意 App 或传感器）攻击感知功能；（3）数据隐私泄露，针对计算和存储功能的数据隐私攻击。针对这些问题，报告介绍了当前研究前沿的技术探索，包括利用区块链技术构建去中心化的安全数据共享机制、应用主观逻辑建立动态可信管理模型、引入前景理论指导资源受限下的最优安全决策。这些技术共同构成了一个旨在提升移动群智感知网络安全性的架构。展望未来，报告强调了三个关键研究方向以提升广义具身智能系统的安全性：（1）发展多模态安全感知能力，利用多源信息融合检测异常与威胁；（2）建立动态信任与信誉机制，有效管理节点间的交互风险；（3）研发自适应防御机制，使系统能够根据环境和威胁变化实时调整安全策略。

广州大学教授李默涵从工业控制系统的视角切入，对具身智能安全问题进行了系统性的层次化分析。报告提出了一种将工业场景下具身智能安全风险划分为宏观（企业级）、中观（车间级）、微观（机器人级）三个层面的研究框架，为理解复杂工业系统中的安全问题提供了结构化视角。这一框架揭示了不同层级面临的关键科学挑战。（1）微观层面：核心问题在于单个机器人本体的访问控制与任务执行安全，面临非授权访问、中间人攻击（MITM）导致的任务中断或危险执行等风险；（2）中观层面：核心问题在于生产网络内部的安全防护与流程协同，面临内网渗透攻击导致的生产调度失当、工艺流程异常等风险；（3）宏观层面：核心问题在于企业级服务接口的安全与供应链韧性，面临对外暴露服务被利用导致错误决策，甚至整个生产或服务链条被破坏的风险。这种分层分析方法不仅清晰地界定了不同粒度的安全问题，也为未来研究跨层级风险传播机制以及设计多层次、协同化的安全防护策略指明了方向。

浙江大学教授冀晓宇的报告深入探讨了具身智能安全相较于传统 AI 安全和大模型安全的本质演变与内涵外延的扩展。报告明确指出了具身智能面临的核心挑战之一是跨域安全威胁，这是一个由信息物理系统特性引发的新型安全问题。其根源在于数字域与物理域之间映射失配导致的带外脆弱性，这种脆弱性可被利用，导致具身智能体出现“精神致幻”（错误的决策或认知）和“肉身失控”（物理行为异常或危险）等严重后果。报告进一步阐释了具身智能跨域攻击的典型范式：攻击者通过构造物理攻击信号（如特定的声、光、电磁信号），以传感器、执行器、电源等物理“肉身”组件为攻击入口，并协同利用具身智能算法层面的脆弱性，最终对智能体自身及其所处物理环境造成安全威胁。针对此类攻击，报告提出了一种前瞻性的防御理念：“先天基因编辑”优于“后天注入疫苗”，强调在传感器等物理组件的设计阶段就嵌入安全与隐私保护能力，这是硬件层面安全研究的前沿方向。同时，具身智能安全需要覆盖全生命周期的评估与防护体系，并亟须建立相应的安全标准、评测工具与平台，这是推动具身智能安全技术成熟和应用落地的基础性工作。

山东大学教授胡鹏飞的报告重点聚焦于具身智能的感知安全风险这一关键领域。报告指出，感知系统作为具身智能体与物理世界交互的入口，其安全性至关重要，但同时面临决策干扰和信息泄露等多样化威胁。报告深入分析了具身智能感知系统易受攻击的特性：这些攻击往往具有物理性、隐蔽性和多样性，能够绕过传统的基于网络的或纯数字域的安全防御措施，对其检测和防御构成了严峻的科学挑战。因此，有效的防御措施必须采取多模态、跨物理域的综合策略，例如整合声学、光学、电磁乃至电力层面的防御手段。未来将有三个重点研究方向：（1）系统性地开发新型具身智能攻击方案，以更全面地发掘和理解潜在脆弱性；（2）深入探索具身智能感知系统的安全边界，明确其在不同环境

和攻击条件下的可靠运行范围；（3）积极融合其他专业领域的知识（如物理学、信号处理、材料科学等），以开发出更具韧性和适应性的高级防御能力。这些方向共同构成了提升具身智能感知鲁棒性和安全性的核心研究议题。

具身智能物理实体正在逐步成为现实，各种人型机器人和无人驾驶汽车作为“肉身”百花齐放。而相应的物理实体问题也随之显现，如感知不可靠、计算不可靠、执行不可靠等。与会专家就信息物理系统融合下具身智能的安全问题进行了激烈讨论，并达成如下共识。

物理攻击与信息攻击的结合可能会带来严重的系统性威胁。具身智能系统不仅会面对传统的信息攻击，如数据篡改或网络攻击；还可能受到物理攻击，如传感器感知攻击、执行器攻击、电源攻击等。当攻击从物理域进入数字域时，可能导致信息层面的错误决策，从而对具身智能体本身以及环境造成危害。这种“物理域－信息域”跨域攻击的联动效应会导致更加严重的系统性威胁。

具身智能“安全执行边界”的定义需要综合考虑其复杂的内部结构和外部环境。在内部结构上，具身智能在设计时要考虑具备对环境的动态感知能力，设计能够在高风险环境下进行自我保护和功能降级的安全策略；在外部环境上，不仅要考虑环境中出现的干扰信号，还要考虑恶意攻击者的行为与攻击能力。

“大脑”与“肉身”应该相互配合实现真正的“端到端安全”。在具身智能中，大脑和肉身之间的紧密协调对于实现“端到端安全”至关重要。为了实现“端到端安全”，具身智能体应该建立主动性与适应性，能够在环境发生变化时迅速响应，并适应新环境中的潜在威胁。信息物理域都应该设置相应的防护措施，在信息域采用鲁棒性算法，在物理域实现健全的传感器和执行器保护。

专题四：具身智能多模态及对齐安全

随着多模态大模型取得突破，不同模态的数据可被对齐到文本特征空间，使得大模型能够综合处理和理解图像、音频以及文本等多模态信息。多模态大模型赋能具身智能，大大增强了其跨模态穿透与融合感知能力，赋予具身智能体“眼”“耳”“口”等五官感知，以及多模态推理能力。虽然新模态的引进增强了具身智能的感知决策能力，但也引入了新的攻击面。例如，攻击者可以在视觉上构造恶意对抗样本，更容易实现越狱攻击。此外，具身智能的行为是否与人类的价值观对齐也是重要的安全问题，特别是在复杂的物理和社会环境中，要防止具身智能体产生恶意行为，保障人类的安全。与会的专家学者围绕这些问题进入了深入讨论，分享了各自的研究成果、实践经验和观点看法。

清华大学副研究员吕勇强的报告探讨了构建可信具身智能系统的核心挑战与实现路径。报告认为，确保系统的安全性是一个涉及硬件与软件协同的根本性科学问题。在硬件层面，研究前沿在于探索利用可信计算技术，如可信芯片及其提供的安全加固机制，来构建安全可靠的底层计算平台，为上层软件和算法提供信任根。在软件与系统设计层面，报告强调了人因工程在具身智能安全设计中的关键作用，这是另一个重要的研究方向。具体的技术挑战包括设计直观且安全的人机交互界面，以及研发鲁棒的异常检测与安全处理机制。未来的研究目标是构建软硬件一体、可形式化验证的安全架构，并发展一套以人为中心的设计原则与方法论，确保具身智能系统在复杂多变的环境下能够稳定运行，并满足人类的安全与信任需求。

北京大学助理教授董豪的报告聚焦于提升具身智能操作的可解释性这一关键科学问题。报告介绍了一项重要的技术前沿：利用物理引擎仿真结合扫描的三维物体模型与真实图像，生成大规模、高质量的混合现实数据集。这种方法旨在解决机器人操作中从模拟到现实的迁移难题，通过在仿真环境中进行充分训练，实现模型在真实物理世界中高度泛化的抓取能力。此外，报告探讨了利用多模态大语言模型提升操作可解释性的潜力，提出通过让大模型生成描述操作策略的自然语言文本，可以作为一种有效的解释机制，增加人类对机器人行为的理解和信任。报告还分享了在具身智能主动探索学习方

面的最新进展，通过离线或在线的探索性尝试来微调模型，旨在发展出高效且在决策逻辑上更具可解释性和可靠性的具身操作能力。这些研究共同推动具身智能从“黑箱”操作向更透明、可信赖的智能行为迈进。

清华大学副教授李琦的报告系统性地剖析了具身智能大模型系统在安全防御层面面临的多重挑战。报告将安全风险归纳为横跨网络环境、代码逻辑、模型决策以及输出价值四个层面的复杂科学问题，为理解和应对此类系统的安全风险提供了整体性框架。网络环境安全的核心挑战在于防御针对系统接口（如 API）的网络攻击和保障运行环境的可信性；代码逻辑安全面临软件供应链攻击、代码漏洞利用等威胁，需要确保系统软件本身的鲁棒性；模型决策可靠性的关键问题在于抑制模型的“事实性幻觉”，确保其决策基于准确信息；输出价值对齐的核心在于防止模型生成或执行违背人类价值观和社会规范的行为。针对这些挑战，报告提出了相应的技术前沿与防御对策，包括：面向 API 安全的网络流量行为分析、针对系统多触发路径的代码漏洞深度检测、基于模型内部状态（如激活值）的事实错误预判，以及运用大模型安全对齐技术进行价值取向纠正。这些防御措施共同构成了保障具身智能大模型系统端到端安全性的多维度研究方向。

北京大学助理教授杨耀东的报告从内生安全与外生安全两个独特视角，深入探讨了具身多模态大模型安全的内在动因与外在挑战。这为理解和评估具身智能安全提供了新的理论维度。报告将内生安全定义为具身大模型在现实决策中自发遵循人类价值观的能力，其核心科学问题是如何实现和验证这种内在的价值对齐。外生安全则关注模型在面对外部干扰时的鲁棒性，包括抵御对抗性攻击和应对分布外场景导致的系统失效，其核心科学问题是如何提升系统在开放、不确定环境下的韧性。报告分享了针对这两类安全问题的评估方法学研究以及潜在的解决方案探索。一个关键的技术前沿与独特挑战在于，具身大模型不仅需要像传统大模型那样与数据和人类偏好对齐，更需要实现与客观物理反馈和物理规律的对齐。报告还结合国内外人工智能立法与治理的现状，强调了从技术和社会层面共同推进具身智能安全研究的紧迫性与重要性，未来的研究须致力于开发能够同时确保价值对齐与物理世界鲁棒性的理论、方法与评估体系。

在集中讨论环节，与会专家对具身智能多模态及对齐安全面临的挑战和已有实践经验进行了充分探讨与交流。大家一致认为，多模态数据处理、模型行为对齐、安全风险分析与治理是提高具身智能系统安全性的重要手段，并形成了如下共识：

- **多模态是具身智能的重要能力。**区别于信息域的视觉语言模型，具身智能部署在物理现实中，不仅需要视觉和文本模态，还需要听觉、触觉等重要模态乃至全模态的感知能力。同时，在引入新模态提升系统决策能力时，还须关注新模态会扩大系统攻击面的安全风险，要在感知前后做好相应的安全防护。
- **对齐是具身智能安全的关键部分。**设计部署具身智能时必须符合人类价值观或是社会公共利益，即与人类价值观对齐。同时，具身智能的决策还须满足现实中客观的物理学规律或限制，即与物理反馈对齐。上述二者应该相互验证、相互补充，从而提升具身智能系统的安全性。

专题五：具身智能网络通信及数据安全

具身智能网络通信及数据安全，指终端在交互感知中收集隐私数据，并与云端交换时需确保数据在处理、存储、传输各环节的安全和隐私。因此，既要保证全流程中隐私数据的安全，又要满足具身智能体的智能化操作。与会的专家学者围绕上述问题进入了深入讨论，分享了各自的研究成果、实践经验和观点看法。

北京航空航天大学教授罗洪斌的报告从网络安全的根源出发，聚焦于网络体系结构创新这一核心科学问题。报告指出，当前 TCP/IP 体系结构在设计之初就缺乏内生的安全考量，导致其存在固有的安全脆弱性，难以应对日益复杂的网络攻击。下一代网络体系结构的研发应借鉴基础科学中的对应原理和对称性原则等思想，探索能够在兼容现有网络

生态的基础上实现安全增强的创新路径。这构成了如何在保障兼容性的前提下进行颠覆性安全创新的关键技术挑战。此外，报告强调了开放系统对于提升网络可管理性、规律性与效率的重要性，暗示了未来安全网络架构设计中，需要在封闭的安全机制与系统开放性之间寻求有效平衡。这也是网络体系结构研究中一个值得深入探讨的议题，其最终目标是构建既安全可控又不牺牲现有网络价值的新一代网络基础架构。

武汉大学教授王骞的报告聚焦于大模型驱动下具身智能的数据安全与隐私保护这一新兴交叉领域。报告指出，虽然大模型极大地挖掘了具身智能在自动驾驶、机器人等领域的应用潜力，但也引入了严峻的数据安全与隐私挑战，特别是大模型自身的安全性问题亟待解决。报告探讨了当前研究前沿的几个关键方向。（1）面向具身场景的数据保护策略：如何针对性地设计数据加密、脱敏、访问控制等机制；（2）训练数据遗忘处理：研究如何在遵守隐私法规的同时，有效且可验证地从大模型中移除特定训练数据，其中结构化遗忘机制是一个重要的技术探索方向；（3）对抗攻击防护：开发能够抵御针对具身智能系统数据输入或模型本身的对抗性攻击的技术；（4）隐私增强学习：探索将差分隐私、联邦学习等技术应用于具身智能大模型训练的可行性与效率。报告呼吁学术界与工业界共同攻关安全评估基准的建立和模型可解释性的提升等难题，这些是保障大模型驱动的具身智能技术健康、可持续发展的关键科学问题。

南京大学教授田臣的报告从更宏观和深刻的视角，重新审视了具身智能系统的整体安全问题，并将其提升到关乎人类社会未来的高度。报告的核心关切在于：随着具备物理行动能力的人工智能体的发展，人类可能不再是地球上唯一的智慧主导力量，这将引发一系列根本性的科学和社会问题。报告援引了杰弗里·辛顿（Geoffrey Hinton）等科学家的观点，警示了强人工智能可能带来的颠覆性社会冲击，包括大规模结构性失业、极端的财富集中，乃至对人类生存构成潜在威胁。报告强调，当前的研究亟须超越单个系统的技术安全，深入探索智能系统与人类社会整体的协同安全。这要求建立新的研究范式，从微观（单个智能体行为）和宏观（社会结构、经济模式、伦理规范）两个层面系统性地研究和应对 AI 发展可能引发的深远社会变革。这代表了 AI 安全研究领域中一个极具前瞻性但也异常复杂的科学挑战。

湖南大学教授靳文强的报告聚焦于物联网（IoT）环境下的感知安全与隐私保护，这与具身智能系统紧密相关，因为后者常依赖于广泛部署的 IoT 设备进行环境感知。随着智能终端设备的激增，如何在不侵犯用户隐私的前提下，安全、有效地采集和处理海量的感知信息，已成为一个关键的技术挑战和科学问题。报告分享了其团队在该领域的研究前沿成果，例如探索实现隐私保护下的高效感知信息处理方法。此外，报告深入剖析了智能终端设备在音频和数据输入方面的具体安全隐患，揭示了新的攻击向量，例如：（1）利用耳机等外设进行音频窃听（如通过分析振动信号）；（2）通过传感器数据（如运动传感器）推断屏幕输入内容，从而泄露敏感信息。针对这些具体的传感器层面的侧信道攻击，报告提出了相应的防御措施。未来的研究方向包括系统性地识别和评估 IoT 设备中的新型感知层漏洞，并开发轻量级、高效的嵌入式防御机制。

山东大学教授成秀珍的报告着眼于具身智能系统中的可验证安全约束问题，并探讨了先进密码学技术作为解决方案的潜力。报告提出的核心科学问题是如何在分布式、可能互不信任的具身智能系统组件之间，在保护数据隐私的同时，强制执行并验证安全相关的计算或策略。报告重点介绍了两种前沿密码学技术：安全多方计算（SMC）和零知识证明（ZKP）。安全多方计算允许多个参与方协同计算一个函数，而无须暴露各自的私有输入；零知识证明则允许一方（证明者）向另一方（验证者）证明某个论断为真，而无须透露任何额外信息。报告指出，这些技术能够实现隐私保护下的数据计算与结果验证，构成了实现可验证多方安全的关键技术支撑。未来的研究前沿在于设计针对具身智能特定场景（如协同感知、分布式控制、安全策略合规性验证）的高效、实用的 SMC 和 ZKP 协议，并将其集成到

实际系统中，以提供强形式化的隐私与安全保障。

在讨论环节，与会专家就具身智能网络通信及数据安全问题展开激烈讨论与交流，并形成如下共识。

• **安全和异构融合是具身智能通信的重要需求。**具身智能在边缘端侧的多样化，以及具身智能系统形态的异构化，使通信协议的需求发生变化，对通信协议的物理层和链路层乃至网络层可能提出新的要求。同时，还须确保具身智能通信数据的安全与隐私保护。

• **辩证地看待低时延需求。**绝对的低时延是一个伪命题，应该辩证地看待具身智能低时延需求。对于高动态场景例如自动驾驶等，则须符合低时延设计或者本地计算能力。而对于通信要求不高的具身智能场景，则还须考虑系统的综合成本来实现网络通信协议的设计。

共识与倡议

共识

1. 具身智能作为迈向通用人工智能最有可能的路径，将大模型技术与机器人实体相融合，是智能科学发展的新范式。这种人机物充分融合形成的复杂系统，正是新质生产力的典型代表，也是科技创新的必由之路。

2. 具身智能作为信息域和物理域中的大模型和机器人技术的融合，是一个横跨物理域和信息域的复杂系统，其必然面临如软硬件协同等更为复杂的安全风险与挑战。具身智能系统一旦失控将造成严重后果，因此具身智能安全的研究需要尽快开展并制定相应的评测标准。

3. 与离身智能相比，具身智能具有与物理世界相交互的特性。相较于过去的网络安全研究，具身智能安全中的许多安全问题将从信息空间延展到物理空间。因此，具身智能安全研究须特别关注信息物理融合的安全问题。

4. 具身智能安全须注重具身智能体与人的交互安全问题，应研究相应的防护技术，避免人身受到损害。

5. 具身智能安全是全流程全生命周期的安全，应该注重软硬件协同防护，需要不同安全研究方向的科研工作者合作。同时，随着具身智能技术的革新，必将出现如具身智能安全责任归属、法律法规监管、伦理等方面新的社会问题，因此具身智能安全也需要学科交叉的合作研究。

倡议

1. 要建立健全全流程下的具身智能安全体系框架，关注传统安全问题在具身智能场景下的变化，并发掘具身智能中独特的安全挑战，有组织地开展具身智能安全研究。同时，须综合考虑学科交叉需求和实际适用性，建立具身智能安全标准与评测平台，形成制式化的涵盖全流程全生命周期的具身智能评测规范。

2. 为保障具身智能安全，须推动具身智能关键软硬件国产化、自主化，打造自主可控的产业链供应链，保障我国具身智能产业安全，提升国际话语权。

3. 推动具身智能安全的产、学、教深度融合，充分发挥学术界与工业界的优势互补，并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，形成人才驱动与技术创新并进的良好局面，推动我国具身智能产业的良性、健康发展。 ■

整理：
陆炫存　黄郑献　郑一凡　张世琦　应天棋
王　禹　李鑫宇
特邀嘉宾：
郭世泽　金耀初　陶建华　蒋树强
参会嘉宾（按姓氏拼音排序）：
成秀珍　董　豪　何银军　胡鹏飞　冀晓宇
金意儿　靳文强　李默涵　李　琦　李树涛
罗洪斌　吕勇强　沈　超　苏　洲　田　臣
王　骞　王　震　王志波　吴秉哲　杨耀东
杨永耀
秀湖会议学术委员会（AC）主席：
胡事民
会议执行主席：
徐文渊　朱浩瑾
会议工作人员：陈国兴