

ISC:以“四大安全智能体”对抗 AI 安全风险

孙庆阳

8月6日,以“ALL IN AGENT”为主题的第十三届互联网安全大会(ISC.AI 2025)在北京开幕。与会嘉宾共同探讨人工智能(AI)加速融入产业核心领域的安全挑战问题并达成共识:“安全是人工智能发展的底线,技术创新与协同共治是保障 AI 安全的核心路径。”

以模制模应对 AI 安全高风险

ISC 名誉主席邬贺铨指出:“百年未有之大变局加速演进时代,以大模型为代表的人工智能技术正在推动全球数字化加速发展,为千行百业的转型升级注入强大动能。”

随着 AI 技术广泛应用,安全风险日益复杂多变。360 数字安全集团总裁胡振泉直奔要点:“AI 安全进入‘高风险期’,内忧外患亟待破局。”他进一步阐释 AI 存在的“四大基因缺陷”,以及外部黑产利用 AI 制造攻击武器,使得传统安全防御体系难以应对。为此,360 创新提出“以模制模”新范式,利用“四大安全智能体”(内容安全智能体、AI Agent 安全智能体、软件安全智能体、安全风险评估智能体)构建全链路防护体系。

国际数据中心(IDC)中国副总裁钟振山表示,安全问题已成为全球企业应用大模型的关键挑战,需“以模型防护模型”,而大模型在安全运营、数据安全等领域的赋能将重塑安



第十三届互联网安全大会纳米 AI 体验区

全市场格局。

邬贺铨还强调了技术创新与架构演进的重要性:“未来的方向在于模块化与解耦化,将庞大的单体模型拆解为灵活的功能组件,通过‘IaaS+PaaS+SaaS’的创新组合,构建新型 AI 平台。”这一思路不仅降低了中小企业应用 AI 的门槛,还显著提升了模型效率与可及性,为安全技术的普及与深化应用奠定了基础。

智能体赋能行业安全生态共建

俄罗斯工程院外籍院士、中国苏州大学机电学院院长孙立宁指出:“在机器人从‘功能

执行’迈向具身智能的时代,机器人成为大模型的‘物理化身’,安全风险从虚拟渗透至现实。”这一变化要求产业界重新审视安全策略,构建从“杀毒软件”到“免疫系统”的全面防御体系。

华为 ICT Marketing 与解决方案销售副总裁史振钰表示:“当前 AI 正处于从‘答得妙’迈向‘靠得住’的关键转折节点,要共建韧性智能体生态,共筑可持续、可信赖的 AI 生产力未来。”

广州汽车集团股份有限公司副总经理閻先庆,以智能网联汽车为例补充道:“安全是智能网联汽车发展的基石,也是必须

坚守的底线,汽车安全生态建设需要全行业共同努力。”

产业协同与生态共建也成为大会共识。全国工商联党组成员、副主席罗来君建议:“广大民营企业要秉持开放普惠原则,加强与高校、科研机构的协同创新,促进跨区域、跨领域的协作,共享技术成果与资源。”

构建智能向善的 AI 安全治理体系

第十三届全国人大社会建设委员会副主任委员、国家互联网信息办公室原副主任、世界互联网大会秘书长任贤良表示:“人工智能之于网络安全,既是风险也是机遇,要在充分发挥人工智能

潜力的同时,确保其发展的安全性、可靠性和公平性。”他呼吁加强全球协作,推动共治共建,坚持智能向善,以普惠促安全。

邬贺铨进一步指出:“面对日益严峻复杂的威胁态势,传统被动、碎片化的安全防护已力不从心,亟需根本性变革。”他提出构建全生命周期的体系化安全运营框架。罗来君对此补充道,“民营企业应在政府指导下参与构建涵盖技术研发、应用落地、安全防护的人工智能全链条的治理框架,打造一批具有引领性、示范性、国际化的标准规范。”

在全球共治方面,任贤良表示:“世界互联网大会国际组织愿与国际社会携手共筑网络空间安全屏障,促进网络安全技术与人工智能融合发展。”这一立场体现了中国在全球网络安全治理中的积极态度与责任担当。通过深化国际合作,共建开放、包容、安全的数字未来,智能向善与全球共治的安全治理体系正逐步形成。

据悉,本次大会期间,360 集团展示了包括“四大安全智能体”在内的大模型安全卫士,发布了全球首个 L4 级别的企业智能体工厂——SEAF,全面覆盖 AI 内容安全、智能体执行安全、供应链安全、风险评估及企业知识管理与智能协作等领域。这意味着智能体在安全防护与产业落地层面已从理论构想迈向实际应用阶段。

具身智能机器人十大发展趋势发布

本报讯(记者 李洋)《2025 具身智能机器人十大发展趋势》8 月 8 日在 2025 世界机器人大会上发布。

十大发展趋势分别是:

第一,物理实践、物理模拟器与世界模型协同驱动的具身感知。物理实践是具身智能的本质,物理模拟器可以构建高保真的训练环境,世界模型可以提供环境当中比较本质的内部特征。三者融合既可以保证丰富有效、真实的环境,也可以用于训练具身智能机器人与环境的接触和非接触交互的感知能力,为其决策和控制奠定基础。

第二,多层次端到端的具身决策。由多模态大模型启发的具有数理基础的认知与规划研究,与生命科学家的成果

融合,并与实时的控制模块融合,可以显著增强具身智能机器人在非结构化环境下的泛化性和实用性。

第三,从控制角度来看,可以融合模型预测、强化学习和生命科学的具身智能控制。一方面可以把模型预测控制的动态优化能力,把强化学习自适应决策融合起来,更进一步地与生命科学的冗余多环路控制机制相融合。这样可以更加让具身智能机器人向人发展,实现具身智能的新控制,提升其在新环境当中的适应性和高性能。

第四,生成式人工智能驱动的具身智能机器人设计。通过对于电机、减速器、驱动器、结构、连接件和材料的统一优化,同时与工材领域的科学成果相结合,在物理模拟器当中

实现硬件与控制策略的协同优化,可自动探索任务中实现最优的具身智能的机器人设计。

第五,高度协同与动态适配的具身智能软硬件一致性。具身智能机器人需要软硬件的一致性,在硬件开发的阶段需预置适配算法的接口规范,在算法的设计当中又会内嵌物理约束,就是软中有硬,硬中有软,并且通过联合仿真验证,就是在有软有硬的情况下,让系统更加保持一致,让软件模块更加接近硬件,让整体系统更加符合我们的软硬件一致性的期望。

第六,具身智能机器人工厂,在仿真环境当中实现自然语言交互、环境生成、机器人本体设计、决策-控制算法以及软硬件一致性算法等研发,让

它们有机地结合在一起,并且反复进化。这样的系统可以根据性能和需求实现快速设计,实现高质量具身智能机器人系统,为社会服务。

第七,具身智能大规模高质量数据集,基于物理实体采集与仿真合成构建大规模高质量具身智能数据集。这里高质量是一个关键,关于大规模,科研的期望是让规模变小。同时可以显著提升具身智能机器人的本体构型优化、多模态训练效率及跨场景策略迁移能力。

第八,具身智能机器人集群及与人协同的发展,融合多智能体的协同机制,构建具身智能机器人集群。同时不断提升具身智能机器人的安全性,及其与人的共情能力,让具身智能机器人真正走向人类,成

为人类的朋友。

第九,跨学科的具身智能机器人开源社区。首先具身智能机器人的发展需要信息科学、工程与材料科学、数学物理科学、生命科学等多学科协作,将在全球范围内聚集各领域的顶级科学家和工程人员,促进具身智能领域的技术探讨,助力产业链的上下游深度融合和协作发展。

第十,面向具身智能机器人的安全评估与伦理建设,通过行为规范验证、决策可解释性分析和数据安全性研究等,能够确保建立面向具身智能机器人的安全评估体系和伦理规范。确保在复杂开放环境中决策的可靠性、可解释性以及行为的安全性,这才使得具身智能机器人能够走向服务行业。