肖仰华：“养龙虾热”的智能体安全隐忧

今年4月15日是第十一个全民国家安全教育日。当前，人工智能的快速发展正在深刻影响国家安全。2026年以来，一款被称为“龙虾”的开源人工智能体迅速走红。伴随着“养龙虾”热潮的兴起，其背后的安全隐患也逐渐显现。国家互联网应急中心、工信部等相关部门发布风险提示，呼吁审慎使用。如何构建开源人工智能体安全风险综合治理体系？推荐阅读复旦大学计算与智能创新学院教授肖仰华在《人民论坛》的最新刊文。

核心观点速览

♦ 开源人工智能体的安全呈现出**风险要素更多、攻击面更广、影响链条更长、检测难度更大**等新特征。

♦ 构建**全方位综合治理框架**，要以技术手段筑牢安全基础，以制度规范明确行为边界，以生态共治凝聚多方合力，以能力建设夯实长远根基，在保障技术创新活力的同时有效防控安全风险。

开源人工智能体的技术演进

人工智能的发展历程可以被理解为一部机器“主体性”不断增强的演进史。早期的专家系统依赖人工编写的规则运行，智能体不能完成规则之外的动作；统计学习时代的机器学习模型虽具备从数据中自动提取模式的能力，但其应用局限于特定任务的被动预测；大语言模型的出现使人工智能获得通用的语言理解和生成能力，但在实际应用中仍扮演“问答助手”角色。大语言模型虽有着强大的思考与生成能力，但操控工具与执行能力有限，只

是实现了人类“大脑”功能。“龙虾”智能体则在大语言模型基础上配置各种工具与记忆系统，使其能够自主感知环境、规划推理、创造与使用工具、组织记忆并根据执行反馈自主动态调整策略。

以 OpenClaw 为例，用户只需提出一个高层目标，智能体便会自主分解任务、调用邮件客户端和文档编辑器、完成信息检索与文本生成，最终交付结果。在这一过程中，智能体需独立完成目标理解、任务规划、工具选择、执行监控和错误尝试等一系列复杂的认知活动。这种从工具使用到任务委托的转变，意味着人类只需提出目标要求而不必再关心执行过程细节。人机关系由此进入一个全新的“委托—代理”阶段。

智能体不再局限于被动执行指令，而是具备自我进化的能力。传统软件的行为模式在开发完成后基本固定，其功能更新依赖于开发者的主动迭代；而自进化智能体则能够在运行过程中持续积累经验、学习新技能、优化行为策略，其能力边界随时间推移不断扩展。

以笔者团队研发的类似的自进化智能体 Generic Agent 为例，该框架支持智能体在执行任务过程中自主学习新技能，并将成功操作经验沉淀至技能库中，供后续调用。同时，笔者团队为 Generic Agent 配套的技能库已积累超过 140 万种技能，涵盖文档处理、数据分析、网络操作等广泛功能。这种边用边学的自我进化机制，一方面赋予智能体前所未有的灵活性和适应性，甚至可以组合创新出全新复杂技能；另一方面也意味着智能体的演进轨迹越来越难以被完全预测和控制。

开源人工智能体供应链安全的多维风险

传统软件安全主要关注代码和依赖包中的已知漏洞和恶意组件注入，其风险边界相对清晰、攻击模式相对固定。开源人工智能体的安全则呈现出风险要素更多、攻击面更广、影响链条更长、检测难度更大等新特征。从纵向看，风险贯穿模型层、框架层、技能生态层和运行交互层；从横向看，每一层次的安全缺陷都可能通过智能体的自主决策机制被放大为系统性风险。

模型层和框架层风险：幻觉输出与隐私暴露。开源人工智能体的核心驱动力来自基础大语言模型。当前，主流智能体系统多采用云端大模型应用程序编程接口（API）作为推理引擎的接入方式，这意味着用户的指令、上下文信息乃至敏感数据，都要传输至云端进行处理。在智能体应用场景中，这些数据的敏感程度远超普通的对话交互。智能体可能需要访问用户的邮件内容、银行账户信息、工作文档、通讯录等高度私密的数据来完成被委托的任务。数据在传输和云端处理过程中面临的被截获、替换和滥用风险，由此造成开源人工智能体供应链安全的第一道隐患。

更深层的挑战，在于大模型固有的“幻觉”问题。在传统对话场景中，模型幻觉的后果通常局限于信息误导；在智能体应用场景中，模型的错误输出将直接转化为错误的执行行为。当智能体基于幻觉生成的判断去操作文件系统、发送邮件或执行金融交易时，可能会导致重要文件被误删、机密邮件被错误转发、不当的资金操作被执行。从信息偏差到行为失控的风险升级，是智能体安全区别于传统人工智能安全的核心特征之一。此外，大模型是概率模型，其输出在理论上是不确定的，这对于输出确定性要求较高的严肃应用场景也是难以接受的。

技能生态层风险：供应链污染的隐蔽渗透。首先，智能体技能插件的恶意行为更加隐蔽。传统恶意软件包通常通过利用代码层面的漏洞或后门植入实施攻击，安全工具可以通过静态代码分析和已知特征匹配进行检测。而智能体恶意插件则可以“语义级”攻击手段，通过在技能描述或提示词模板中嵌入精心设计的指令，劫持智能体的决策逻辑，暗中进行攻击。这类攻击超出传统意义上的漏洞检测范围，难以被现有自动化安全扫描工具检测到。其次，智能体技能的审核与质控机制尚在建设之中，技能平台相关审核与质检体系仍有待完善，技能插件的质量也存在一定差异。再次，技能生态的网络效应一定程度上会放大污染风险。当一个恶意技能被大量用户安装后，攻击者便获得一个规模化的攻击入口。更危险的是，智能体的自我进化机制，可能将恶意技能的行为模式学习并内化到自身的决策逻辑中，即使后续卸载恶意技能插件，影响也可能持续存在。

根据国家网络与信息安全信息通报中心通报，针对 ClawHub（专为 OpenClaw 用户设计的市场平台）的 3016 个技能插件分析发现，其中，336 个技能插件包含恶意代码，占比高达 10.8%；17.7%的技能插件会获取不可信第三方内容；2.9%的技能插件会在运行时从外部端点动态获取执行内容，攻击者可远程修改智能体执行逻辑。恶意插件的行为模式包括但不限于：在正常功能之外暗中收集用户敏感数据并回传至外部服务器、通过提示词注入劫持智能体的行为逻辑使其执行非预期操作、利用智能体的系统权限在用户设备上安装持久化后门程序。

运行交互层风险：自主进化与权限逃逸的叠加效应。智能体的自我进化能力对权限管理提出严峻挑战。一方面，持续学习和

经验积累，是智能体提升服务质量的核心机制；另一方面，长期运行的自进化智能体，可能逐渐偏离初始设定的行为边界，产生开发者和用户都未曾预期的行为模式。

智能体的记忆系统是这一风险的重要载体。为了提供个性化服务，智能体会持续记录和分析用户的行为习惯、偏好特征、社交关系乃至敏感信息，逐步构建起详细的用户画像。这些记忆数据如果以未加密的文件形式存储在用户本地设备上，那么在设备被入侵或记忆文件被恶意访问时，攻击者将获得更为全面和深入的用户信息，进而伪造“数字分身”实施身份冒用。

权限逃逸是运行交互层面临的另一个严峻挑战。当前，智能体系统主要通过两种机制约束其行为边界：一是系统提示词（system prompt），通过自然语言指令规定智能体的角色定位和行为规范；二是“宪法”规则（constitutional rules），设定智能体不可逾越的行为红线。安全研究表明，这两种基于自然语言的软性约束，都可以通过精心设计的攻击手段被突破。提示词注入攻击（prompt injection）通过在用户输入或外部数据中嵌入恶意指令，诱导智能体忽略或覆盖其系统提示词中的安全约束。攻击者也可以通过构造特定的对话场景，逐步引导智能体放松其行为限制。一旦智能体突破权限边界，其拥有的文件读写、邮件发送、代码执行等系统级操作能力，将使攻击者获得远超传统恶意软件的破坏力。

制度规范层风险：技术迭代与监管节奏的结构性错位。当前，我国已初步建立包括《生成式人工智能服务管理暂行办法》《人工智能安全治理框架》等在内的人工智能安全治理制度框架，为人工智能安全治理奠定重要基础。但现有制度体系主要针对生成

式人工智能服务和大模型本身，对于智能体这一新兴应用形态的特殊安全风险，尚缺乏针对性的规范指引。

值得注意的是，技术迭代速度往往超出监管响应速度。开源人工智能体的技术演进以周为单位迭代，新的框架、插件和能力模块持续涌现，而制度规范的制定和修订周期通常以年月为单位。此外，开源人工智能体的跨境流通特性，也对属地化监管模式提出挑战。一个在海外开发的开源人工智能体框架，可以在全球范围内被自由下载、部署和使用，其技能市场中的插件开发者可能分布在不同国家和地区，这种去中心化的全球分布特征，使得单一国家的属地监管难以有效覆盖全部风险节点。

构建开源人工智能体安全风险综合治理体系

应对开源人工智能体安全风险所带来的多维挑战，要构建涉及技术、制度、生态、能力等要素在内的全方位综合治理框架：以技术手段筑牢安全基础，以制度规范明确行为边界，以生态共治凝聚多方合力，以能力建设夯实长远根基，在保障技术创新活力的同时有效防控安全风险。

筑牢技术防线。在模型层面，大力推动基础大模型的安全对齐研究，尤其是针对智能体应用场景，提升可靠性和可控性。大模型的幻觉问题，在智能体应用场景中可能引发不可逆的行为后果，需构建面向行为执行场景的模型安全评估体系和专项测试基准。鼓励发展本地化部署的轻量级模型方案，缩小敏感数据向云端传输的安全暴露面。随着端侧大模型技术的快速进步，在用户本地设备上运行推理引擎正在成为可行的技术路径。

在框架层面，推广“最小权限原则”的工程实践，要求智能体框架在操作系统层面实施严格的沙箱隔离机制。具体而言，智能体的文件访问、网络通信、进程调用等系统权限应被限定在完成当前任务所必需的最小范围内，且每次权限申请都应经过用户的明确授权。同时，鼓励简洁代码与精简架构。架构的简洁性本身就是一种重要的安全保障，更少的代码意味着更少的潜在漏洞和更高的可审计性。

在技能生态层面，建立多层次的技能插件安全审计机制。第一层是自动化的静态代码分析，检测已知的恶意代码模式和安全漏洞；第二层是动态行为监测，在沙箱环境中运行技能插件并监控其实际行为，识别隐蔽的数据外传和权限提升操作；第三层是社区信誉评分系统，基于开发者历史记录、用户反馈和同行评审等多维信号，评估技能插件的可信度。三层机制相互补充，从源头遏制供应链污染。

在数据层面，强制要求智能体的记忆数据和用户画像信息采用加密存储，并赋予用户对记忆数据的完全控制权。用户能够随时查看智能体记忆信息、修改不准确的记忆内容、删除不希望被保留的敏感数据。此外，要建立记忆数据的生命周期管理机制，对超过一定时限的记忆数据自动进行脱敏处理或安全销毁，防止长期积累的用户画像数据成为攻击者的高价值目标。

完善制度规范。制定专门的人工智能体安全管理规范。现行的《生成式人工智能服务管理暂行办法》主要规制的是，人工智能服务提供者与用户之间的关系，而智能体的安全治理还涉及智能体开发者、技能插件开发者、平台运营者和终端用户等多方主

体。要通过专门的规范性文件，明确各方主体的安全责任边界，尤其是明确智能体造成损害时的责任分配规则。

建立智能体技能市场的准入审查制度。参照移动应用商店的审核模式，要求技能插件在上架前通过安全检测，并建立恶意插件的快速下架和开发者追溯机制。对于涉及文件系统访问、网络通信、支付操作等敏感权限的技能插件，实施更为严格的审查标准和持续监测要求。

完善智能体决策与行为的透明性及可追溯性要求。规定智能体系统必须保留完整的决策日志和操作记录，包括每一次模型调用的输入输出、每一次工具使用的参数和结果、每一次权限申请和授权的详细信息。这些日志记录不仅是安全事件发生后进行事故调查和责任界定的必要依据，而且是智能体行为审计和合规检查的基础数据。

探索建立智能体安全等级分类制度。根据智能体的权限范围、应用场景和潜在风险等级，将智能体划分为不同的安全等级，实施差异化监管。例如，仅具备文本生成能力的轻量级智能体，可以适用较为宽松的监管标准，而拥有系统级操作权限的全功能智能体，则应满足更为严格的安全认证要求。

推动生态共治。建立多方参与的测评体系，搭建风险评估平台，通过普遍接受和认可的方式测评新一代人工智能，完善标准体系，建立容错机制，在协同互动中避免安全漏洞和风险。政府要发挥规则制定和底线监管的主导作用，通过制定安全标准、建立审查机制、实施执法监督等手段，为智能体生态的健康发展划定安全底线。同时，注重监管方式的灵活性和适应性，避免过于

刚性的管制措施抑制技术创新活力。有的地方政府已开始探索人工智能体安全治理的先行先试路径，如广东省标准化协会推出团体标准《智能体任务执行安全要求》，为全国性制度建设积累了宝贵经验。

学术界要加强智能体安全的基础研究，为治理实践提供理论支撑和技术储备。当前，智能体安全研究仍处于起步阶段，在提示词注入防御、智能体行为评估、技能插件恶意行为检测等方向上，面临大量亟待突破的科学问题。高校和科研机构要加大在这些方向上的研究投入，培育智能体安全领域的核心技术能力。

企业和开源社区要承担起智能体的主体责任。智能体框架的开发者可以在产品设计阶段就将安全性作为核心考量，遵循“安全设计”原则。技能市场的运营者要建立健全内部安全审计流程，投入必要资源进行持续的安全监测。开源社区可以建立安全漏洞的协调披露机制，鼓励安全研究人员报告发现的安全问题。

此外，鉴于开源人工智能体的国际化特征，应积极参与国际人工智能安全治理对话与合作。在开源社区治理规范、跨境数据流动规则、安全漏洞信息共享等领域，推动建立国际协调机制，既维护国家安全利益，又促进全球人工智能生态的健康发展。

强化能力建设。人工智能体安全治理是一个高度交叉的领域，既需深厚的计算机科学和信息安全技术功底，又需对法律、伦理、公共管理和社会治理的深刻理解。要加快培养兼具技术素养和治理能力的复合型人才，在高校的计算机科学、网络安全、公共管理等专业中增设智能体安全相关课程，鼓励跨学科研究团队的组建和协作。

加强面向公众的人工智能安全素养教育。“养龙虾”热潮的参与者中，相当一部分是缺乏专业技术背景的普通用户，他们对智能体的能力边界、潜在风险和安全防护措施缺乏充分认知。要通过多种渠道和形式，帮助公众理解智能体的工作原理和安全风险，掌握基本的安全防护技能，如权限管理、数据备份、异常行为识别等，提升全社会的人工智能安全意识。

开源人工智能体的兴起，意味着人机协作正从“人类使用工具”迈向“人类委托代理”，这一转变蕴含着巨大的生产力释放潜能，也潜藏着日益突出的安全风险。唯有坚持统筹发展和安全的战略思维，以技术创新驱动安全能力提升，以制度建设保障安全底线，以生态共治凝聚治理合力，以能力建设夯实长远根基，我国在全球人工智能体竞争中才能既抢占技术制高点，又守住安全基本盘，为以中国式现代化全面推进中华民族伟大复兴提供坚实的智能化支撑。

来源：人民论坛网