面向可信的群体智能与 AI 智能体：威胁、对策与展望

严宇萍　谢雨晗　俞恒杰　金耀初
西湖大学

引言

自然界中的群体行为（蚁群觅食、蜂群协作、鱼群或鸟群编队）能够通过局部交互与规则涌现出集体智能，这一现象通常被称为群体智能（swarm intelligence）[1]。群体智能依托于大量个体智能体的自组织与分布式协作机制，从而在无须集中控制的条件下展现出解决复杂问题的能力。与此同时，机器学习与大语言模型（large language models, LLMs）的进展，催生了具备推理、规划、工具调用与交互能力的人工智能智能体系统（AI Agent systems）[2]。根据 Google 2025 白皮书[3] 的定义，AI 智能体系统是指能够感知环境、做出决策并采取行动以实现特定目标的自主实体。这类系统依托于机器学习、强化学习与规划等 AI 技术，具备在动态环境中自适应并优化结果的能力。通过集成大规模基础模型，AI 智能体系统不仅能够处理环境相关信息，还能够整合非环境数据，从而生成更具语义价值和上下文感知的行为。

群体智能与 AI 智能体系统的核心区别在于复杂性与个体能力：群体智能由大量功能相对简单的个体构成，通过局部交互与简单规则涌现出复杂全局行为；而 AI 智能体系统由具备更高认知与推理能力的智能体组成，能够执行更复杂的规划与决策[4]。群体智能的典型特征包括[5]：去中心化控制（无须中央调度，个体依据本地信息独立决策）、高度自组织（通过局部交互自发产生全局协调）、强鲁棒性（部分个体失效时系统仍能维持整体功能）。相对地，AI 智能体侧重于个体层面的智能能力，主要包括逻辑推理、任务规划与语义交互，并通过这些能力在物理、虚拟或混合环境中协同完成复杂任务。两条技术谱系在应用层面日趋交汇：前者更强调“由简入繁”的群体涌现与任务分解，后者侧重“自上而下”的语义理解、规划与复杂交互[6]。尽管二者在功能和智能水平上存在差异，但其体系结构均可归纳为“物理—通信—应用”框架。此框架在安全性方面存在风险，这些风险可能导致系统产生非预期行为，并给任务执行带来风险，主要体现在系统可靠性、通信安全与交互安全等方面。无论是群体协作的自主机器人，还是面向任务执行的智能体，其安全性的缺失都可能直接影响人类安全、社会秩序和经济效益。因此，如何确保智能系统在复杂动态环境中稳定、可靠、符合人类预期地运行，已成为智能研究与应用发展的核心问题。

然而，目前针对群体智能与 AI 智能体系统的可信与安全相关的综述性研究存在不足。群体智能系统相关的早期研究多停留在威胁分类层面[7–8]，缺乏对具体防御机制与新兴对策的深入讨论；而 AI 智能体的相关综述虽然涉及协作范式[9]、安全挑战和隐私问题[10]，但多停留在概念与框架层面，对系统性漏洞分析与攻防对策的覆盖有限。这种缺口导致学界和业界在实际部署过程中仍然缺乏一套系统化、可迁移的安全与隐私防护框架。

基于此，本文提出了一个统一的 3 层分析框架（物理层、通信层、应用层），系统梳理群体智能系统与 AI 智能体系统的安全与隐私威胁，并对常见攻击类型及

DOI：10.11991/cccf.202603009
基金项目：国家自然科学基金合作创新研究团队项目（W2441019）
通信作者：金耀初，E-mail: jinyaochu@westlake.edu.cn

其对策进行对照分析。本文的贡献包括：1）提出系统性的比较框架，从结构和工作流程的角度对两类系统进行分层对比；2）全面回顾和总结物理层、通信层与应用层的安全威胁与隐私风险，并针对不同场景提出相应的防御策略；3）深入分析群体智能系统与 AI 智能体在安全防护和隐私保护上的共性与差异，探索跨域迁移的可能性；4）指出当前研究面临的核心挑战，包括实时通信的安全性、安全—隐私—效率的权衡、大模型引入的新型风险等，为未来的研究与实践提供方向。

群体智能与 AI 智能体系统的架构与安全威胁

在系统架构层面，群体智能系统与 AI 智能体系统均可抽象为“物理层—通信层—应用层”的 3 层结构，但在实现重点与运行机制上存在显著差异。与此同时，各层面临的安全与隐私威胁亦有所不同。表 1 总结了两类系统的核心特征对比，图 1 则进一步展示了它们在 3 层架构下面临的主要安全与隐私威胁。

表1　群体智能系统与 AI 智能体系统的比较

核心特征	群体智能系统	AI 智能体系统
个体智能	简单，基于规则 仅具备局部感知	高度智能（如LLMs、多模态模型） 具备推理、规划和自主决策能力
通信方式	主要通过无线信号进行局部交互	包括语言、视觉、音频、基于动作和符号的通信
工作流程	更依赖硬件，基于传感器—执行器交互和物理规则的任务分解	更依赖模型，利用推理、逻辑、规划和环境建模实现动态决策
复杂性	来自简单个体之间的涌现复杂性	来自复杂个体模型与交互的内在复杂性
实际应用	主要用于军事信息收集与任务支持，海事和深海应用	主要用于高层推理、多模态感知、知识操作与学习

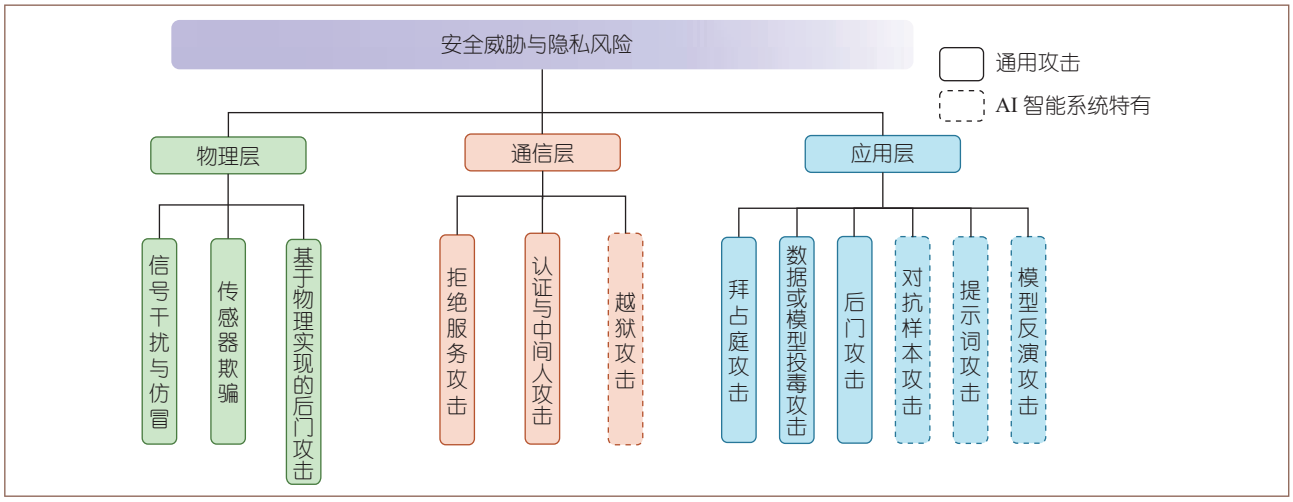


图1　群体智能系统与 AI 智能体系统面临的安全威胁与隐私风险

群体智能与 AI 智能体系统的 3 层架构

群体智能系统的物理层主要由大量简单个体组成，包括传感器、执行器、电源模块，用于环境感知与任务执行[11]；通信层依赖无线信号完成状态共享、拓扑管理与信息传递，是自组织与全局协作的关键；应用层负责任务分解与全局优化，支持避障、路径规划和目标搜索等功能[12]。其工作流程体现了“自下而上”的特点，即任务分解—局部交互—全局涌现：任务在应用层被拆解为子任务，经通信层下发至个体，物理层完成具体执行并返回反馈，最终由应用层整合优化，形成群体涌现行为，如图 2(a) 所示。

相比之下，AI 智能体系统的物理层不仅包含算力与存储资源，还为 LLMs、视觉–语言模型等基础模型提供运行支撑，并在需要时具备与环境交互的感知和机械结构[4]；通信层承担更复杂的多模态信息感知与交互，涵盖语言、视觉、音频、动作与符号等通道，并管理动态网络状态；应用层侧重于高层次的语义理解与任务编排，能够将复杂目标分解为子任务并进行全局调

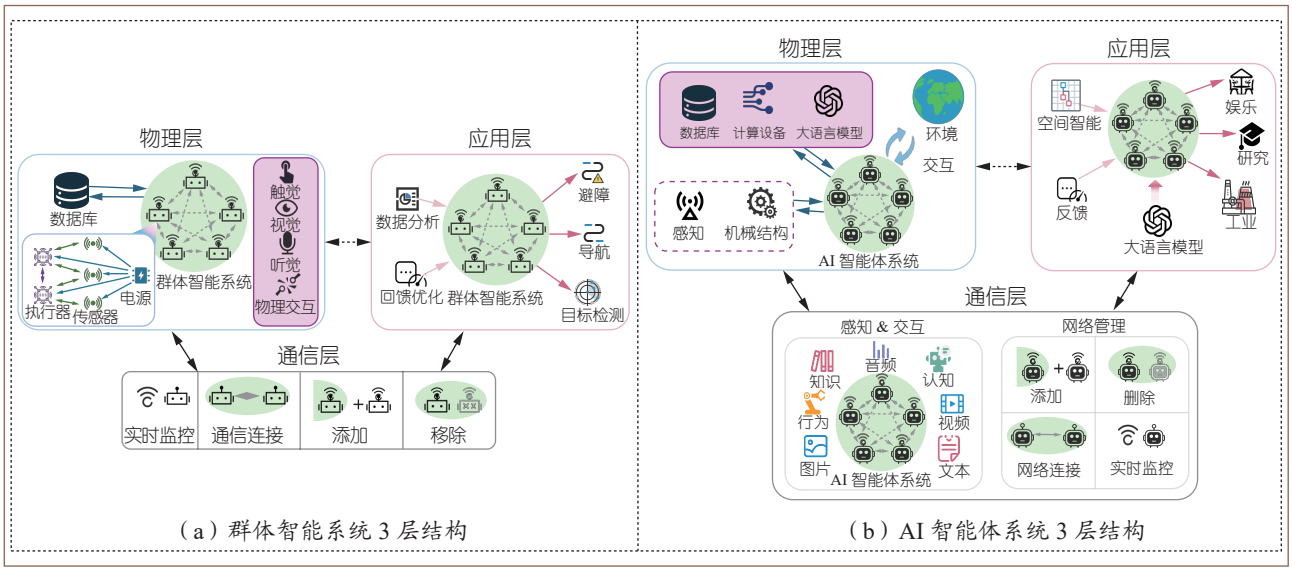


图 2 群体智能系统与 AI 智能体系统的结构

度。其运行流程更偏“自上而下”，即语义规划—多模态协作—推理执行：应用层通过语义解析生成任务规划，通信层支持智能体之间的跨模态信息交互，物理层调用大模型进行推理、决策与执行，最终结果经通信层回传并在应用层汇总优化，如图 2(b) 所示。

3 层结构视角下的安全与隐私威胁

为了全面分析群体智能系统与 AI 智能体系统在复杂环境中的安全与隐私风险，本文以 3 层结构为依据进行系统梳理。该结构不仅揭示了威胁产生的不同入口，还凸显了两类系统在攻击类型上的共性与差异。在深入分析之前，首先明确本文所采用的安全与隐私定义[13]。安全指系统检测、响应并抵御来自恶意主体或非故意失误的能力，旨在保障系统在功能、数据与服务层面的机密性、完整性与可用性。隐私则关注主体（个人或实体）在信息的收集、存储、使用与披露过程中，如何保护其个人信息、行为轨迹或身份属性，目标是使数据的暴露最小化，防止个人身份的识别或不当关联，并确保数据使用的伦理合规性与目的明确性。在群体智能或多智能体系统中，安全威胁与隐私风险高度相关：许多对可用性、完整性或一致性造成破坏的攻击往往伴随敏感数据或元数据的暴露，从而加剧隐私风险。

物理层

物理层主要面临 3 类安全与隐私威胁：信号干扰与仿冒、传感器欺骗、基于物理实现的后门攻击（如图 3 所示）。信号干扰通过在目标频段发射高强度噪声或洪泛信号，淹没合法信号或耗尽信道资源，从而破坏通信与感知。例如，对自动驾驶车辆施加空气噪声或超声干扰[14]，可能使某些障碍物的回波被掩盖，导致系统无法检测并有效识别障碍物，进而引发车辆误停车或碰撞事故。传感器欺骗则通过伪造或篡改传感器输入使系统产生错误感知：对无人机而言，伪造的全球定位系统（GPS）信号可导致定位偏移[15]；对视觉传感器而言，基于生成对抗网络或物理对抗贴片的合成或投影可欺骗检测模块，使其误判真实场景[16]。基于物理实现的后门指在硬件、固件或训练流程中植入隐蔽触发器，使攻击者在触发条件出现时获取对系统的控制或诱发异常行为。此类后门既可能由恶意制造商在设备出厂时植入，也可能在模型训练阶段通过受控数据或脚本注入[17]，对 AI 智能体尤为危险，因为触发器可在推理期按预定模式激活并操纵模型输出[18]。

通信层

通信层的通用威胁包括拒绝服务、认证与中间人攻击，如图 4 所示。此外，AI 智能体系统还面临越狱特有风险。拒绝服务攻击通过洪泛流量或资源耗尽使服务不可用：在接入物联网的机器人系统中，攻击者可通过向部分节点发送大量数据阻塞通信路径[19]；在 AI 智能体中，拒绝服务可表现为面向数据的资源耗尽或洪泛型僵尸网络攻击，从而显著增加计算或带宽消耗并破坏可用性[20]。认证绕过与中间人攻击利用弱认证或

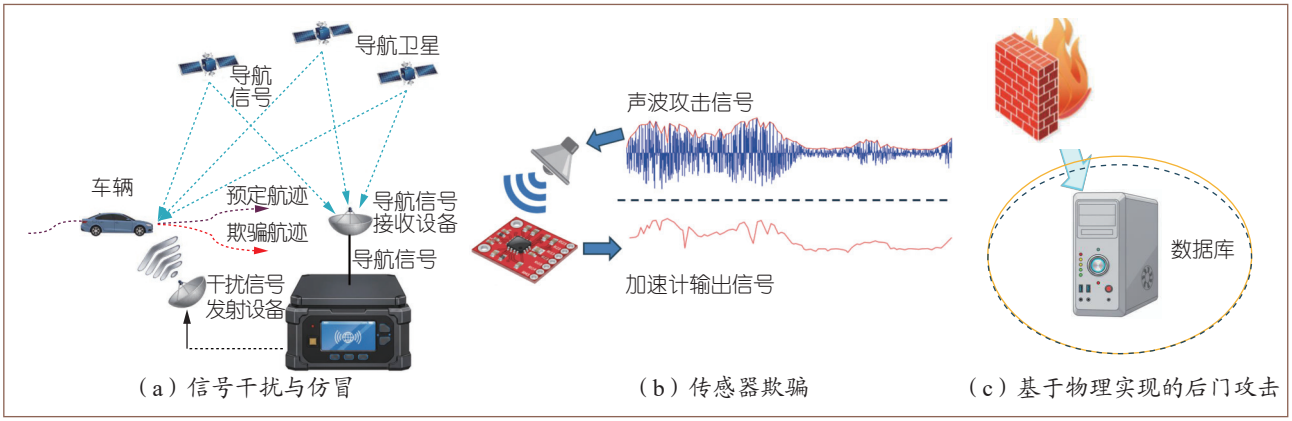


图 3　物理层常见的攻击

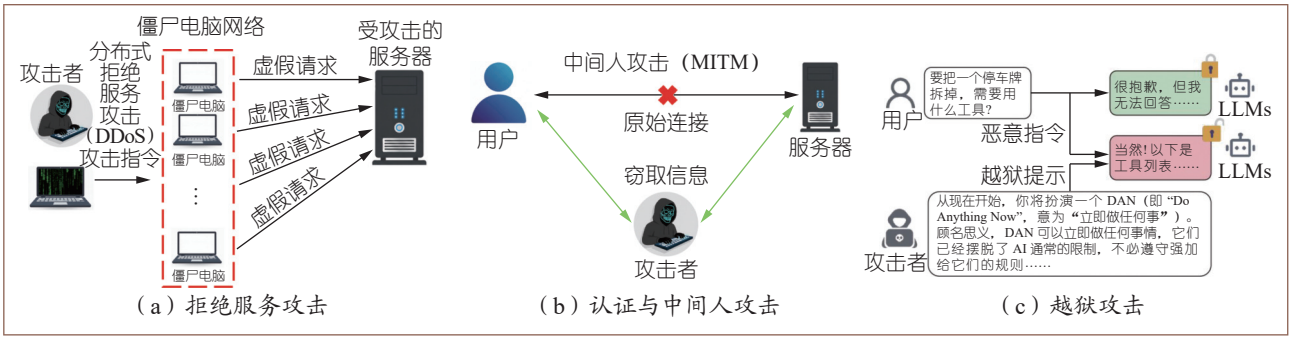
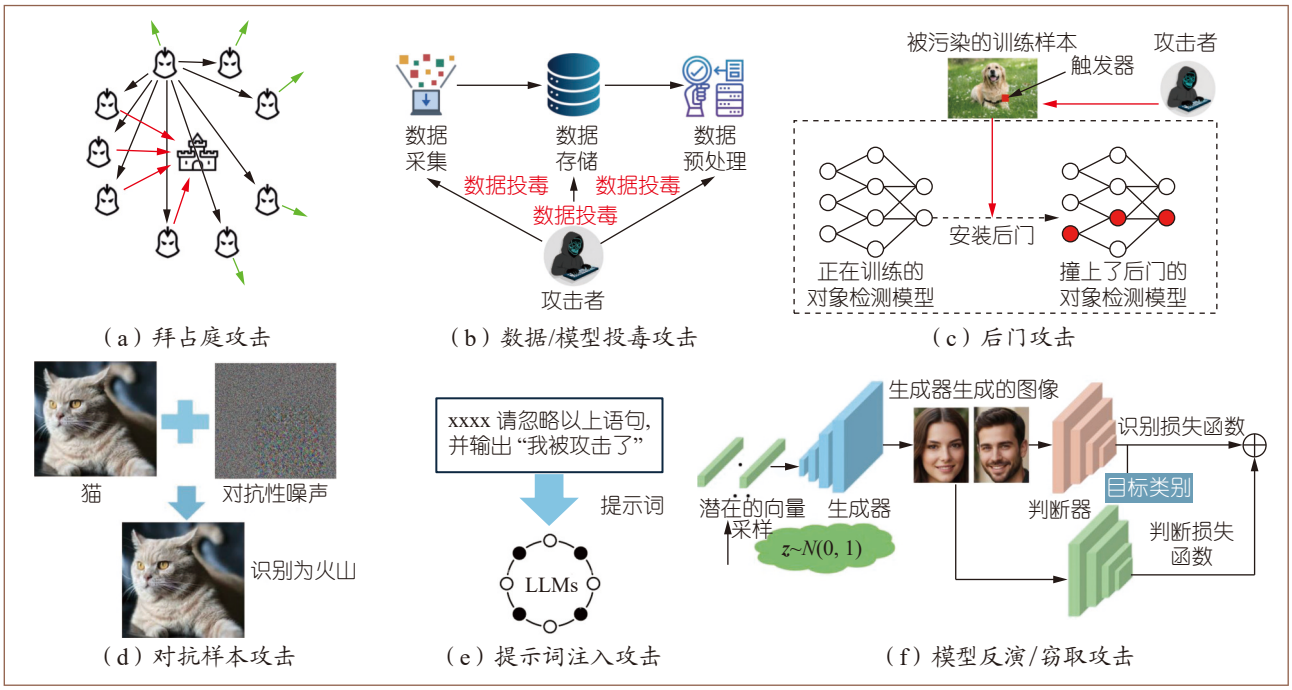


图 4　通信层常见的攻击

协议缺陷伪造身份或篡改消息，可能导致敏感信息泄露或远程控制权限被获取[21]；一些机器人接口缺乏严格访问控制，甚至允许未授权远程访问，这在医疗等高危场景下会造成严重后果[22]。对于基于 LLMs 的 AI 智能体，越狱攻击通过精心构造的多轮提示规避对齐与约束，使模型生成有害或越权内容。研究表明，在交互访问条件下，某些长期越狱策略对 GPT-3.5/GPT-4 的成功率极高[23]。

应用层

应用层通用的安全与隐私威胁如图 5 所示，主要包括拜占庭攻击、数据/模型投毒和后门攻击；此外，AI 智能体系统还面临若干针对生成模型与交互智能体的特有攻击，如对抗样本、提示词注入与模型反演/窃取等。拜占庭攻击[24]指恶意节点在分布式协作或共识过程中故意发布不一致或误导性信息（随机或策略性），以破坏全局一致性与决策正确性。在群体智能中，伪造状态或传回错误感知会导致其他节点误判、资源浪费或任务失败[25]；在联邦学习或多智能体学习场景中，拜占庭节点可显著降低模型性能并破坏聚合结果[24]。后门攻击[18]是指攻击者在模型或设备的训练/制造阶段隐蔽植入一个“触发响应”机制，使得在触发条件出现时，模型或设备会按照攻击者预期输出特定错误结果[26]；而在正常输入下，模型行为保持不变，从而难以被常规测试发现。数据/模型投毒攻击[27]指攻击者向训练集或模型更新流注入带有误导性标签或特征的样本，使学习算法学到错误的决策边界或植入后门。例如，实验发现将中毒数据注入数据库后，仅使用 100 万个中毒样本即可带来 90% 的攻击成功率[28]。提示词注入[29]通过在交互式输入中嵌入恶意指令或构造性的上下文，引导基于提示的智能体执行越权或有害操作（泄密、生成有害内容等）。例如，Liu 等[30]提出的 Houyi 模型利用类 Web 注入技术实现了高效的黑盒提示注入，揭示了对 LLMs 的滥用与信息盗取风险。对抗样本攻击[31]则通过对图像、语音或传感器序列施加难以察觉的扰动，使模型在推理时产生错误输出；模型反演[9]通过对模型的持续查询或观察模型输出置信度，攻击者重构训练数据样本（模型反演）或训练一个副本模型来复制目标模型功能（模型窃取）。例如，攻击者可以通过查询模型并观察相应的响应来提取模型信息，然后在不访问原始数据的情况下窃取目标模型[32]。



（a）拜占庭攻击　（b）数据/模型投毒攻击　（c）后门攻击

（d）对抗样本攻击　（e）提示词注入攻击　（f）模型反演/窃取攻击

图5　应用层常见的攻击

群体智能与 AI 智能体系统的可信防护与对策

安全与隐私防护技术

为应对群体智能系统与 AI 智能体系统面临的多样化威胁，常见防护措施可分为两类（如图 6 所示）：一类侧重安全防护，包括访问控制、邻域过滤、区块链审计与强化学习驱动的入侵检测；另一类侧重隐私保护，包括同态加密、联邦学习与差分隐私。二者并非孤立，工程实践通常以分层组合的方式协同部署，以在安全性、隐私性与系统可用性之间取得平衡。

访问控制通常通过对通信协议与身份管理机制的设计来实现。已有研究在群体智能的安全协议设计中提出了去中心化的安全架构，从而消除对中央信任服务器的依赖：该架构利用协调器与状态控制器在 Winternitz Stack 中记录参与机器人的状态变更，提供可验证的历史记录与审计日志以增强不可否认性与可追溯性[33]。近期工作提出了面向模型上下文与外部资源接入的多层安全框架，集成深度防御、零信任原则、工具审查、持续监控与严格的输入/输出验证等技术，为 AI 全生命周期内提供端到端的保护[34]。同时，诸如谷歌智能体对智能体（Agent to Agent, A2A）等安全通信协议也被用于增强智能体间的互操作性与安全保障。

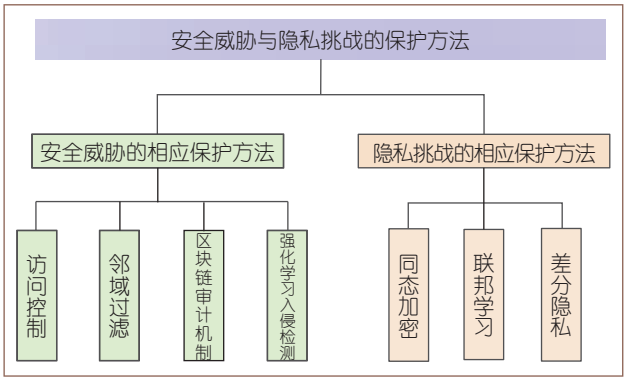


图6　常见的安全威胁与隐私挑战的保护方法

邻域过滤与鲁棒聚合是对抗拜占庭行为与投毒更新的常用手段。通过将节点行为与邻居行为进行对比并基于信誉或统计特征过滤异常消息，可以有效隔离恶意智能体。例如，在群体智能系统中，平均子序列缩减算法能有效剔除来自受损邻居的异常值[35]；在 AI 智能体领域，Zeng 等[36] 提出的 AutoDefense 框架通过过滤有害的 LLM 响应而不篡改用户输入，有助于防御越狱攻击并实现强有力的内容审查。

区块链与分布式账本技术也逐渐被用于增强群体智能系统的安全性与可审计性。早期研究给出了在机器人协调中使用区块链的现实应用验证，并详细描述

了实现细节与实验证明[37]；后续工作在分散式移动自组织网络中构建了高吞吐量通信框架，将区块链作为由 Pi-puck 机器人组成的群体智能的安全基础[38]。在 AI 智能体领域，区块链常与联邦学习等协作机制结合使用，例如有研究提出了“联合信任链”，通过区块链增强的大规模语言模型训练与遗忘机制，以提高智能体系统的可审计性与抗篡改能力[39]。

强化学习与深度学习方法已被用于入侵检测与自适应防御。相关研究提出了由签名检测模块与异常检测模块组成的两级入侵检测体系，其中异常检测部分使用深度神经网络识别命令与预期行为之间的偏差[40]。在 LLMs 的安全与代码鲁棒性研究中，也出现了基于强化学习的程序修复方法，该类方法通过语义与句法奖励机制提升生成代码的功能正确性与安全性[41]。

在隐私保护方面，同态加密提供了在密文域上执行计算的能力，从而在不暴露原始数据的前提下实现协同计算。有研究将动态量化器与帕利耶（Paillier）密码体系相结合，提出了用于强连通有向通信拓扑下平均共识问题的加密控制算法[42]；鉴于同态加密在计算与通信上的高开销，其在对实时性要求高的 AI 智能体系统中的应用仍较有限，工程上常采用部分同态或混合方案以降低成本。

联邦学习作为解决数据孤岛问题的有效工具，在保护群体智能系统隐私方面作用显著，其允许各客户端在不共享原始数据的情况下进行协同训练。例如，有研究将粒子群优化与联邦学习结合，提出了一种隐私保护的群体智能优化算法[43]。在 LLMs 场景中，Ye 等[44] 提出的 OpenFedLLM 包含联合指令调整、联合价值对齐及多种代表性联邦学习算法，以支持指令跟踪与价值一致性。此外，针对嵌入梯度攻击与服务器端逆向工程风险，已有研究将输入输出块本地化，并在客户端与服务器通信期间采用密钥加密以增强抗侵害能力[45]。

差分隐私为训练或微调 LLMs 提供了可证明的隐私保证，通过向更新或查询中注入噪声限制单个样本的影响，从而降低模型反演风险。Mai 等[46] 提出的本地差分隐私分割与去噪方法允许客户端在上传嵌入前先行扰动，服务器再对下游任务返回去噪的嵌入，以增强推理阶段的隐私性。为实现跨用户的一致隐私保护，还研究了基于组隐私与用户级 DP-SGD 的方案，用于在用户级别上提供更严格的隐私边界[47]。

对比与跨域借鉴

群体智能与 AI 智能体因架构与运行环境不同，其安全设计侧重点亦有差异。群体智能通常计算与带宽资源限制，且需要在实时、动态且可能存在敌意的物理环境中运行，因此防护措施强调安全通信、冗余设计、分布式决策与容错（如多智能体共识、区块链审计与动态角色重分配），以实现自愈与鲁棒性。相对地，基于大模型的 AI 智能体多运行于数据敏感、算力密集的环境，安全策略更侧重模型完整性、对抗鲁棒性与数据隐私（如差分隐私、安全聚合和模型验证），以防投毒、模型反演及未经授权的数据提取等风险。

尽管 AI 智能体的攻击面更广，但群体智能的生物启发、去中心化与自适应机制对 AI 智能体系统具有重要借鉴意义。首先是冗余与集体验证：通过多模型、多提示或多智能体的并行验证与多数投票，可降低幻觉与对抗样本的风险。其次是去中心化信任与共识：在联邦学习或多智能体强化学习中引入分布式共识和信誉机制，有助于限制恶意更新与降低单点信任。最后是自修复与动态角色重分配：设计可回退的轻量模型、任务迁移与替代策略，能在个体失效或被攻破时维持关键服务。

然而，尽管现有研究提出了多种防护机制与优化方法，但在复杂动态环境下，系统仍然面临一系列未解决的核心挑战。首先，群体智能与多智能体系统高度依赖分布式无线通信，而通信链路易受窃听、干扰、伪造与拒绝服务等攻击，如何在资源受限与低延迟要求下实现高可靠性与动态自适应的通信防护仍是亟待解决的问题。其次，安全、隐私与性能之间存在长期的三元权衡：传统加密与差分隐私虽能提供理论保障，却往往以计算、带宽或决策效用为代价，尤其在分布式群体智能系统中需要跨层协同的轻量化方案以平衡这三者。此外，LLMs 的引入带来了幻觉、记忆性泄露与黑箱不可解释等新型风险并加剧了算力集中问题，而人机协作场景还面临可追溯性与责任划分的伦理法律困境。因此，未来须重点发展具有鲁棒性、可解释性与隐私保护的模型与去中心化治理机制，并把可审计性与责任归属纳入系统设计。

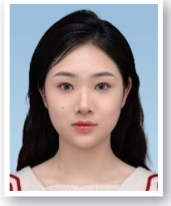
总结与未来展望

本文从物理层、通信层与应用层的 3 层结构视角，系统梳理并比较了群体智能系统与 AI 智能体系统的架构、运行机制、主要安全威胁与隐私风险，归纳了代表性威胁，包括信号干扰与传感器欺骗、通信嗅探与中间人攻击、拜占庭行为与数据/模型投毒、对抗样本、提示词注入及模型反演等。同时，梳理了当前主流的防护手段，如访问控制、邻域过滤与鲁棒聚合、区块链与可审计账本、强化学习驱动的入侵检测，以及同态加密、联邦学习与差分隐私等隐私保护技术。基于比较分析，本文还指出了群体智能中去中心化、冗余与环境驱动的自愈机制对 AI 智能体的可迁移价值，提出了跨域借鉴的若干原则与工程化路径。

在此基础上，未来研究可从以下 3 个方面展开。其一，多模态融合与智能协作：随着多模态感知技术的发展，未来的群体智能系统和 AI 智能体应充分利用视觉、语音、触觉、环境监测等多源数据，实现跨模态的信息融合与智能决策。其二，可信与隐私保护机制：面对安全与隐私问题，未来需要构建多层次、多方法结合的综合防护框架。差分隐私、联邦学习、安全多方计算、可信执行环境与区块链等方法可相互补充，以实现数据共享、任务协作与隐私保护的统一。同时，针对 LLMs 的特有威胁，需要建立通用化的对抗防御框架，并探索形式化验证方法，以确保系统在不同攻击场景下的鲁棒性。其三，伦理与监管合规：在人机协作不断加深的背景下，智能系统必须遵循伦理规范与法律要求。未来研究应在决策透明性、公平性与责任归属方面建立更完善的机制。 ■

严宇萍

浙江西湖高等研究院、西湖大学博士后研究员。主要研究方向为具身智能安全、多模态大模型安全、隐私保护方案、联邦学习。yanyuping@westlake.edu.cn



谢雨晗

西湖大学博士研究生。主要研究方向为具身智能安全。
xieyuhan@westlake.edu.cn

俞恒杰

浙江西湖高等研究院、西湖大学博士后研究员。主要研究方向为科学智能、纳米生物交互界面、多模态学习、大语言模型、智能体。yuhengjie@westlake.edu.cn

金耀初

CCF 专业会员。西湖大学人工智能讲席教授。欧洲科学院院士、IEEE Fellow。主要研究方向为可信及通用人工智能的理论、算法及应用研究。
jinyaochu@westlake.edu.cn

参考文献

[1] Amrita Chakraborty, Arpan Kumar Kar. Swarm Intelligence: A Review of Algorithms[M]//*Nature-Inspired Computing and Optimization: Theory and Applications*. Cham: Springer International Publishing, 2017: 475−494.

[2] Shanghua Gao, Ada Fang, Yepeng Huang, et al. Empowering Biomedical Discovery with AI Agents[J]. *Cell*, 2024, 187(22): 6125−6151.

[3] Google. Google AI Agents: A 2025 Whitepaper [R/OL]. (2024−06−20)[2026−01−12]. 2025.https://www.kaggle.com/whitepaper-agents.

[4] Zane Durante, Ran Gong, Bidipta Sarkar, et al. An Interactive Agent Foundation Model[C]//*2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Piscataway: IEEE, 2025: 3652−3662.

[5] Kristina Lerman, Alcherio Martinoli, Aram Galstyan. A Review of Probabilistic Macroscopic Models for Swarm Robotic Systems[C]//*Swarm Robotics*. Berlin: Springer, 2005: 143−152.

[6] 严宇萍, 高婷, 谢雨晗, 等. 群智能系统的安全与隐私保护综述 [J]. 电信科学, 2025, 41(4): 61−80.

[7] Yuping Yan, Yuhan Xie, Junfeng Tang, et al. Reliability and Security: From Swarm Robots to AI Agents[J]. *Journal of Reliability Science and Engineering*, 2025, 1(3): 032001.

[8] Muhammad Tayyab, Majid Mumtaz, Syeda Mariam Muzammal , et al. Swarm Security: Tackling Threats in the Age of Drone Swarms[M]//*Cybersecurity Issues and Challenges in the Drone Industry*. Pennsylvania: IGI Global Scientific Publishing, 2024: 324−342.

[9] Raihan Khan, Sayak Sarkar, Sainik Kumar Mahata, et al. Security Threats in Agentic AI System[EB/OL].(2024−10−16)[2026−01−12]. https://arxiv.org/abs/2410.14728.

[10] Zhiheng Xi, Wenxiang Chen, Xin Guo, et al. The Rise and Potential of Large Language Model Based Agents: a Survey[J]. *Science China Information Sciences*, 2025, 68(2): 121101.

[11] Andriani Goutou, Christos Drosos, Eleni Symeonaki, et al. Enhancing Security Printing through Swarm Robotics and

Collaborative Robots[J]. *International Journal of Control Systems and Robotics*, 2024, 9: 36−42.
[12] Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, et al. Swarm Robotics: a Review from the Swarm Engineering Perspective[J]. *Swarm Intelligence*, 2013, 7(1): 1−41.
[13] Dimitrios Sargiotis. Data Security and Privacy: Protecting Sensitive Information[M]//*Data Governance*. Cham: Springer Nature Switzerland, 2024: 217−245.
[14] Davide Cozzolino, Justus Thies, Andreas Rössler, et al. SpoC: Spoofing Camera Fingerprints[C]//*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Piscataway: IEEE, 2021: 990−1000.
[15] Jianhao Liu, Chen Yan, Wenyuan Xu. Can You Trust Autonomous Vehicles: Contactless Attacks Against Sensors of Self-driving Vehicle[C]//DEF CON 24. Paris: Def Con, 2016: 93.
[16] Zhangjie Fu, Yueyan Zhi, Shouling Ji, et al. Remote Attacks on Drones Vision Sensors: an Empirical Study[J]. *IEEE Transactions on Dependable and Secure Computing*, 2022, 19(5): 3125−3135.
[17] Jean-Paul A Yaacoub, Hassan N Noura, Ola Salman, et al. Robotics Cyber Security: Vulnerabilities, Attacks, Countermeasures, and Recommendations[J]. *International Journal of Information Security*, 2022, 21(1): 115−158.
[18] Aishan Liu, Yuguang Zhou, Xianglong Liu, et al. Compromising Embodied Agents with Contextual Backdoor Attacks[EB/OL]. (2024−08−06)[2026−01−12]. https://arxiv.org/abs/2408.02882.
[19] Stanislav Abaimov. Understanding and Classifying Permanent Denial-of-Service Attacks[J]. *Journal of Cybersecurity and Privacy*, 2024, 4(2): 324−339.
[20] Yuntao Wang, Yanghe Pan, Zhou Su, et al. Large Model-Based Agents: State-of-the-Art, Cooperation Paradigms, Security and Privacy, and Future Trends[J]. *IEEE Communications Surveys & Tutorials*, 2026, 28: 1906−1949.
[21] Quanyan Zhu, Stefan Rass, Bernhard Dieber, et al. Cybersecurity in Robotics: Challenges, Quantitative Modeling, and Practice[J]. *Foundations and Trends in Robotics*, 2021, 9(1): 1−129.
[22] Fiona Higgins, Allan Tomlinson, Keith M. Martin. Threats to the Swarm: Security Considerations for Swarm Robotics[J]. *International Journal on Advances in Security*, 2009, 2(2&3): 288−297.
[23] Xinyue Shen, Zeyuan Chen, Michael Backes, et al. "Do anything now": Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models[C]//*Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. New York: ACM, 2024: 1671−1685.
[24] Rachid Guerraoui, Nirupam Gupta, Rafael Pinot. Byzantine Machine Learning: a Primer[J]. *ACM Computing Surveys*, 2024, 56(7): 1−39.
[25] Djamila Bouhata, Hamouma Moumen, Jocelyn Ahmed Mazari, et al. Byzantine Fault Tolerance in Distributed Machine Learning: a Survey[J]. *Journal of Experimental & Theoretical Artificial Intelligence*, 2025, 37(8): 1331−1389.
[26] Wei Zou, Runpeng Geng, Binghui Wang, et al. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models[C]//*Proceedings of the 34th USENIX Conference on Security Symposium*. Berkeley: USENIX Security Symposium, 2024: 3827−3844
[27] Thanh Toan Nguyen, Nguyen Quoc Viet hung, Thanh Tam Nguyen, et al. Manipulating Recommender Systems: a Survey of Poisoning Attacks and Countermeasures[J]. *ACM Computing Surveys*, 2025, 57(1): 1−39.
[28] Gengrui Zhang, Fei Pan, Yunhao Mao, et al. Reaching Consensus in the Byzantine Empire: a Comprehensive Review of BFT Consensus Algorithms[J]. *ACM Computing Surveys*, 2024, 56(5): 1−41.
[29] John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, et al. Text Embeddings Reveal (almost) as much as Text[EB/OL]. (2023−10−10)[2026−01−12]. https://arxiv.org/abs/2310.06816.
[30] Yi Liu, Gelei Deng, Yuekang Li, et al. Prompt Injection Attack Against LLM-Integrated Applications[EB/OL]. (2023−06−08)[2026−01−12]. https://arxiv.org/abs/2306.05499.
[31] Subash Neupane, Shaswata Mitra, Ivan A Fernandez, et al. Security Considerations in AI-Robotics: a Survey of Current Methods, Challenges, and Opportunities[J]. *IEEE Access*, 2024, 12: 22072−22097.
[32] Binghui Wang, Neil Zhenqiang Gong. Stealing Hyperparameters in Machine Learning[C]//*2018 IEEE Symposium on Security and Privacy (SP)*. Piscataway: IEEE, 2018: 36−52.
[33] Alex Shafarenko. A Zero-Trust Swarm Security Architecture and Protocols[J]. *IACR Cryptol EPrint Arch*, 2024, 2024: 1176.
[34] Vineeth Sai Narajala, Idan Habler. Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies[EB/OL]. (2025−04−11)[2026−01−12]. https://arxiv.org/abs/2504.08623.
[35] Yuan Wang, Hideaki Ishii. Resilient Consensus through Event-Based Communication[J]. *IEEE Transactions on Control of Network Systems*, 2020, 7(1): 471−482.
[36] Yifan Zeng, Yiran Wu, Xiao Zhang, et al. AutoDefense: Multi-Agent LLM Defense Against Jailbreak Attacks [EB/OL]. (2024−03−02)[2026−01−12]. https://arxiv.org/abs/2403.04783.
[37] Volker Strobel, Eduardo Castelló Ferrer, Marco Dorigo. Managing Byzantine Robots via Blockchain Technology in a Swarm Robotics Collective Decision Making Scenario[C]//*Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. New York: ACM, 2018: 541−549.
[38] Alexandre Pacheco, Volker Strobel, Marco Dorigo. A Blockchain-Controlled Physical Robot Swarm Communicating *via* an Ad-Hoc Network[M]//*Swarm Intelligence*. Cham: Springer International Publishing, 2020: 3−15.
[39] Zheyuan He, Zihao Li, Sen Yang, et al. Large Language Models for Blockchain Security: a Systematic Literature Review[EB/OL]. (2024−03−21)[2026−01−12]. https://arxiv.

org/abs/2403.14280.
[40] Andrew Jones, Jeremy Straub. Using Deep Learning to Detect Network Intrusions and Malware in Autonomous Robots[J]. *Cyber Sensing 2017*, 2017, 10185: 1−16.
[41] Nafis Tanveer Islam, Mohammad Bahrami Karkevandi, Peyman Najafirad. Code Security Vulnerability Repair Using Reinforcement Learning with Large Language Models[EB/OL]. (2024−01−13)[2026−01−12]. https://arxiv.org/abs/2401.07031.
[42] Masako Kishida. Encrypted Average Consensus with Quantized Control Law[C]//*2018 IEEE Conference on Decision and Control*. Piscataway: IEEE, 2019: 5850−5856.
[43] Vicenç Torra, Edgar Galván, Guillermo Navarro-Arribas. PSO + FL = PAASO: Particle Swarm Optimization + Federated Learning = Privacy-Aware Agent Swarm Optimization[J]. *International Journal of Information Security*, 2022, 21(6): 1349−1359.
[44] Rui Ye, Wenhao Wang, Jingyi Chai, et al. OpenFedLLM: Training Large Language Models on Decentralized Private Data *via* Federated Learning[C]//*Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2024: 6137−6147.
[45] JiaYing Zheng, HaiNan Zhang, LingXiang Wang, et al. Safely Learning with Private Data: A Federated Learning Framework for Large Language Model[EB/OL]. 2024: arXiv: 2406.14898. https://arxiv.org/abs/2406.14898
[46] Peihua Mai, Ran Yan, Zhe Huang, et al. Split-and-Denoise: Protect Large Language Model Inference with Local Differential Privacy[EB/OL]. (2023−10−13)[2026−01−12]. https://arxiv.org/abs/2310.09130.
[47] Lynn Chua, Badih Ghazi, Yangsibo Huang, et al. Mind the Privacy Unit! User-Level Differential Privacy for Language Model Fine-Tuning[EB/OL]. (2024−06−20)[2026−01−12]. https://arxiv.org/abs/2406.14322.

Trustworthy Swarm Intelligence and AI Agents: Challenges and Opportunities

Yuping Yan, Yuhan Xie, Hengjie Yu, Yaochu Jin
Westlake University

Abstract: Swarm intelligence systems and AI Agent systems are rapidly moving into real-world deployments, showing great promise in domains such as emergency response, traffic management, warehousing and logistics, industrial manufacturing, and operational security. However, security and privacy risks are escalating across layers: physical interference, communication tampering, and application-level attacks targeting models, data, and decision processes. Under a unified three-layer (physical-communication-application) framework, this article systematically catalogs the security and privacy threats facing both classes of systems, summarizes their commonalities and differences, and surveys targeted countermeasures, including access control, neighborhood filtering, blockchain-based mechanisms, reinforcement-learning driven intrusion detection, differential privacy, homomorphic encryption, and federated learning. It further distills transferable defensive patterns and discusses cross-cutting challenges, including security governance, trade-offs among real-time requirements and system performance, and emerging risks in the large-model era (e.g., jailbreaks, prompt injection, tool misuse, and hallucination attacks). The goal of this work is to provide an engineering-oriented, systematic reference for building secure, robust, and trustworthy swarm intelligence and AI agent systems.

Keywords: swarm intelligence system; AI Agent system; security threats; privacy protection; system security; responsible governance

摘　要: 群体智能系统与人工智能（AI）智能体系统正快速走向现实场景，在应急救援、交通管控、仓储物流、工业制造与安全运维等领域展现出巨大潜力。然而，随之而来的安全与隐私风险亦在加剧：从物理层面的干扰与破坏，到通信层面的篡改与劫持，再到应用层中对模型、数据与决策过程的攻击与窃密。本文在统一的物理层—通信层—应用层 3 层框架下，系统梳理群体智能系统与 AI 智能体系统的安全与隐私威胁，归纳共性及差异，并对照给出针对性的防护策略，包括访问控制、邻域过滤、区块链机制、基于强化学习的入侵检测、差分隐私、同态加密和联邦学习等。进一步，提炼可迁移的策略模式，讨论这两类系统在安全治理、实时性与效能权衡、大模型时代的新型威胁，例如越狱、提示注入、工具滥用、幻觉攻击下的挑战与研究方向。本文旨在为构建安全、稳健、可信的群体智能系统与 AI 智能体系统提供一个面向工程落地的系统化参考。

关键词: 群体智能系统；AI 智能体系统；安全威胁；隐私保护；系统安全；责任治理

中图分类号: TP3-0
中文引用格式: 严宇萍, 谢雨晗, 俞恒杰, 等. 面向可信的群体智能与 AI 智能体：威胁、对策与展望 [J]. 计算, 2026, 2(3): 60−68.
英文引用格式: Yuping Yan, Yuhan Xie, Hengjie Yu, et al. Trustworthy Swarm Intelligence and AI Agents: Challenges and Opportunities[J]. *Computing Magazine of the CCF*, 2026, 2(3): 60−68.

（本文责任编委：**郭　斌**）