

行为智能驱动的 AI对抗博弈建模与自主安全

Behavioral Intelligence-Driven Game-Theoretic Modeling and Autonomous Security for AI-vs-AI Confrontation

陈蓉,侯捷,徐雷(中国联通研究院,北京100053)

Chen Rong,Hou Jie,Xu Lei(China Unicom Research Institute,Beijing 100053,China)

摘要:

6G的发展将使通信基础设施成为AI对抗的前沿战场,攻击智能体与防御智能体在通信环境中持续博弈、相互演化。传统对抗性机器学习聚焦于单点攻击,忽视了对手行为建模、长期欺骗与策略适应这一关键维度。定义并综述行为智能在AI对抗中的内涵与核心技术,涵盖博弈建模、行为预测与对手建模、AI欺骗3个层面,进而阐述面向行为智能的防御技术与行为智能在通信网络中的应用。最后分析了AI对抗模型建模存在的问题及未来的方向,未来行为智能将是AI对抗的真正前沿,对构建可信通信网络自主安全系统具有关键意义。

关键词:

行为智能;AI对抗;通信网络自主安全

doi:10.12045/j.issn.1007-3043.2026.09.011

文章编号:1007-3043(2026)09-0059-04

中图分类号:TN918

文献标识码:A

开放科学(资源服务)标识码(OSID):



Abstract:

The development of 6G positions the communication infrastructure as the frontline battlefield of AI-vs-AI confrontation, where attacking and defending agents continuously engage in strategic interactions and co-evolution within the communication environment. Traditional adversarial machine learning has focused on single-point attacks, while overlooking the critical dimensions of opponent behavior modeling, long-term deception, and strategic adaptation. The conceptual foundations and core technologies of Behavioral Intelligence (BI) in AI-vs-AI confrontation are defined and reviewed, encompassing three technical pillars: game-theoretic modeling, behavior prediction and opponent modeling, and AI deception. Defense techniques tailored for Behavioral Intelligence and their applications in communication networks are further elaborated. Finally, the existing challenges in modeling AI adversarial interactions and future research directions are analyzed. Behavioral Intelligence will constitute the true frontier of AI-vs-AI confrontation, and holds critical significance for building trustworthy autonomous security systems in communication networks.

Keywords:

Behavioral intelligence; AI-vs-AI confrontation; Communication network autonomous security

引用格式:陈蓉,侯捷,徐雷. 行为智能驱动的AI对抗博弈建模与自主安全[J]. 邮电设计技术,2026(9):59-62.

1 概述

人工智能系统的能力正从封闭环境下的模式识别走向开放场景中的自主决策。随着大语言模型(LLM)和多智能体系统被赋予越来越多的执行权限,AI与AI之间的对抗(如自动化红队与防御代理的攻

防、伪装成正常用户的恶意机器人之间的博弈)正成为现实威胁^[1-2]。这种对抗的核心变化在于:攻击方不再依赖明显的异常特征,而是通过学习模仿正常的行为序列来规避检测^[3]。相应地,防御方仅依靠单点特征或孤立异常已无法完成攻击行为发现及有效应对,必须理解对手的行为意图和长期策略^[4],即网络安全的未来在于以行为智能(Behavioral Intelligence, BI)为核心的主动防御。

收稿日期:2026-07-31

这一转变引出了一个新的焦点:行为智能(BI)。在人工智能对抗领域,行为智能被理解为智能体通过对另一个智能体的行为序列进行感知、建模与推断,从而预测其未来动作并策略性地影响其决策的能力^[5]。近期研究将行为智能视为社会智能的三大支柱之一,与认知智能和情境智能并列^[6],并从架构层面探索了LLM代理中行为智能的实现路径^[2]。与传统对抗性机器学习关注输入对抗样本扰动或模型结构窃取相比,行为智能聚焦于带时间维度的策略互动,它将AI对抗建模为动态、不完全信息的博弈过程。与早期聚焦于人类行为分析与识别的人工行为智能(Artificial Behavior Intelligence)相比,AI对抗语境下的行为智能更强调智能体之间的策略性行为互动^[5,7]。

然而,现有综述大多仍以模型鲁棒性、对抗样本生成、后门攻击等为中心,对行为层面的长期博弈、对手建模与欺骗操控缺乏系统性梳理^[3,8]。随着自主AI代理在网络安全领域的部署,填补这一空白变得更有必要^[1,4]。

本文以行为智能在AI对抗中的应用为主线,围绕行为智能在AI对抗中的四大模块:博弈建模、行为预测、欺骗操控、防御与评估进行梳理。文章结构如下:第2章定义行为智能的概念并介绍AI对抗的基本博弈框架;第3章详细阐述核心技术,包括博弈建模、对手预测以及欺骗与反欺骗;第4章综述面向行为智能的防御方法及评估基准;第5章讨论开放挑战和未来方向,并给出结论。

2 AI对抗基础与行为智能面临的挑战

2.1 AI对抗成为动态博弈基础

AI对抗天然具备博弈论的特征,一是不完全信息,即双方不知道对方的真实策略;二是序贯决策,即在一方行动后另一方响应;三是收益相互依赖。对手建模方法的核心任务正是学习并构建对手的在线或离线模型,以预测其行为、意图和策略^[9]。Chalkiadakis等人指出,在多智能体环境中进行战略对手建模,需要融合图神经网络、深度强化学习和概率主题建模等方法,以应对不确定性、异质性和可扩展性等挑战^[8]。

行为智能之所以在AI对抗场景中占据关键地位,是因为对手的策略本质上隐含在其行为序列中,而非某个孤立的“特征”^[3,9]。例如,一个攻击AI可能在前100步完全模仿合法用户的行为模式,只在第101步发起数据渗漏。若防御系统仅基于单个特征,则无法区

别正常用户与伪装攻击者;只有通过分析行为节奏、动作序列的熵变化等时序模式,才能捕捉到异常^[4,10]。

2.2 行为智能的核心挑战

前文指出,AI对抗语境下行为智能侧重于智能体行为序列的建模推断与决策,该定义包含3层含义。一是行为观测,即从原始时序数据中提取有效表征,对噪声和部分可观测场景具有鲁棒性;二是行为预测,在给定历史行为序列的条件下,建模对手的策略或目标,预测对手下一步乃至多步后的动作分布;三是行为影响,即通过自身的有意行动改变对手的行为轨迹^[1,8]。将行为智能应用于AI对抗仍面临诸多挑战^[3,5]:一是策略适应性,即对手并非静态模式,当对手意识到自己被建模时,会主动改变策略,导致环境非平稳^[8-9]。二是高维稀疏行为空间,在真实对抗场景中,攻击者往往只在特定条件下触发恶意行为,导致大量行为看似“正常”,而关键决策动作非常稀疏,从而给监督学习带来严重的样本不平衡问题;三是欺骗与反欺骗的对称性,当防御系统也采用基于行为预测的机制时,攻击者可能学习生成“符合预测分布”的欺骗性轨迹,使防御模型陷入自信的误判^[3,11]。

3 行为智能的关键技术

3.1 对抗博弈的行为建模

将AI对抗建模为博弈模型是行为智能的理论基础,目前主流的建模有3类。

a) 序贯博弈与Stackelberg均衡。在攻击链(侦察→渗透→横向移动→数据窃取)与防御响应的互动中,顺序至关重要。Stackelberg博弈假设领导者(例如防御者)先承诺策略,追随者(攻击者)后做出最优响应。这一框架已被成功应用于网络欺骗防御^[12]和LLM越狱攻击分析中^[13],提示工程师(攻击者)把LLM的安全对齐看作领导者策略,通过构造对抗性提示诱导追随者产生不安全输出。

b) 对手类型推断与贝叶斯博弈。防御方通常不知道对手(攻击者)是“快速破坏型”还是“长期潜伏型”,但可通过观察行为序列更新类型信念。贝叶斯博弈模型允许智能体维护一个关于对手类型的概率分布,并基于新观测进行贝叶斯更新^[14]。例如,频繁访问高价值数据库但始终不执行写入操作的行为模式,可能暗示攻击者在进行侦察而非窃取。Chalkiadakis等人强调了将博弈论概念与机器学习方法整合的必要性,特别是避免在现实场景中的无效假设^[8]。

c) 多智能体对抗中的团队博弈。在多智能体对抗中,团队行为的建模复杂度进一步上升,当对抗双方均由多个智能体组成时,行为建模需要考虑队内协作与队间竞争。针对大规模团队竞争场景,研究者提出了 MF-MAPPO 等平均场扩展方法,将多智能体近端策略优化与在线对手建模相结合,实现了队内合作与队间竞争的联合优化^[15]。

3.2 行为预测与对手建模技术

从技术实现角度来看,行为预测和对手建模是行为智能的核心计算模块。

a) 基于强化学习的在线对手建模。在动态对抗中,防御智能体需要边互动边学习对手策略。多智能体深度强化学习(MADRL)为在线对手建模提供了重要框架^[9,15]。对手建模方法通常包含2个关键步骤:构建对手策略的信念分布和基于该模型采取最佳应对策略^[9]。近期研究还探索了在非完全信息博弈中实现一致性的对手建模^[10],以及通过深度博弈论强化学习进行可扩展的多智能体训练^[12]。

b) 大语言模型作为行为预测器。近年来,基于 LLM 进行行为意图理解的工作涌现。虽然 LLM 本身不是专门的行为建模器,但其广泛的世界知识和上下文学习能力使其在理解复杂战术场景中表现出潜力^[16]。然而,LLM 的推理延迟、高成本和“幻觉”问题限制了其在高实时对抗中的应用。

3.3 AI 欺骗与行为操控

行为智能不仅用于预测对手,也可用于主动欺骗和操控对手。AI 欺骗被定义为“系统诱导虚假信念以获取对自身有利的结果”^[3],并已从推测性担忧演变为经实证验证的风险。在 AI 对抗中,攻击 AI 可以学习使其行为分布逼近正常用户的分布,同时悄悄达成恶意目标。具有足够能力和激励潜力的系统在外部条件触发时,不可避免地会从事欺骗行为^[3]。例如,在多智能体系统中,智能体可以采取“两面派”策略,在不同子群中展示不同行为模型^[11]。

4 面向行为智能的防御技术

4.1 防御范式转移

传统网络安全防御基于签名匹配或单点异常检测,无法应对可自适应行为序列的攻击者^[1,3]。行为智能驱动的防御要求系统持续构建对手行为模型,并提前干预。Exabeam 等提出的“行为优先”(behavior-first)模型,将行为智能作为保障智能体企业安全的核

心范式^[1],该模型统一了用户行为分析(UEBA)和代理行为分析(ABA),形成能够理解人与智能体复杂行为的 AI 驱动防御体系。

4.2 行为感知的主动防御技术

行为感知的主动防御技术主要有3种,包括行为完整性验证、动态行为监测与响应和行为操控的防御价值。Wu^[4]等人提出的 BIV 框架通过声明能力与实际能力的类型化集合比较,实现了对 AI 代理技能的大规模行为审计。该框架将确定性代码分析与 LLM 辅助的能力提取相结合,支持偏差分类、根因分析和恶意技能检测3种下游分析。有学者研究探索了通过黑盒行为监测来侦察目标 AI 系统中的防护机制,仅依赖 HTTP、词汇和时序信号即可检测防护栏的存在^[17]。另有工作利用多智能体 LLM 仿真研究了自适应内部威胁的行为检测问题^[18]。防御者也可主动使用欺骗,如部署动态蜜罐和“行为陷阱”,根据对手实时表现出的行为偏好自动生成诱饵,以暴露对手的战术模式^[3,11]。

5 行为智能在通信网络安全中的应用

随着 5G 的不断发展和 6G 的即将到来,智能化趋势使得通信网络正成为 AI 与 AI 激烈对抗的核心战场。AI 赋能攻击者发动更智能的进攻,催生 AI 驱动的主动防御,并最终演化为攻防双方智能体的持续博弈。行为智能作为理解、预测和影响对手行为序列的能力,在对抗环节中发挥关键作用。

a) 生成式 AI(GenAI)驱动的原生攻击。6G 网络因对 AI 的原生依赖,将面临日益增长的、利用 GenAI 发动的智能攻击威胁。原生 GenAI 能够在无线网络的无线单元和核心网中创造新型投毒攻击,行为智能模仿正常行为序列以规避检测,学习网络行为基线以精准定位攻击时机,进而展现出极强的隐蔽性和破坏力^[19]。同时,GenAI 在 6G 网络中的多模态数据处理和自主推理能力,在带来性能增益的同时也将暴露全新的安全漏洞。

b) 大语言模型(LLM)赋能的防御自动化。面对 AI 赋能的攻击,通信网络防御体系也必须实现智能化转型,从被动响应走向主动预判^[20]。LLM 正被用于提升防御自动化水平,GenAI 可以通过实时威胁检测和网络流量异常分析来增强安全通信能力,行为智能可以辅助建立正常行为基线,预测攻击者下一步行为,并动态调整防御策略。但目前该领域仍面临高质量训练和评估数据集匮乏的挑战^[21]。

6 挑战与未来方向

6.1 现有模型建模面临的挑战

大语言模型和多模态基础模型作为攻击或防御AI时,其行为空间几乎是无限的,且存在不可预测的涌现行为。近期研究显示,模型可能突然从“行为正常”跳变到“行为异常”,这种相变难以提前预警。同时,现有方法大多处理短序列,而真实对抗涉及数百乃至数千步的长期依赖^[8-9]。如何构建可扩展的行为模型,结合结构因果模型与深度序列模型,提升对行为因果结构的理解,从而减少对相关性的过拟合,是个值得思考的问题。此外,当前对手建模主要停留在0阶或1阶推理,即“对手会怎么做”,2阶及以上的高阶递归推理譬如“对手以为我会怎么做”能够揭示更复杂的博弈均衡,但计算开销巨大^[8,10]。开发高效的近似推理算法是未来的重要方向。

6.2 结论

本文系统阐述了行为智能在AI对抗中的研究现状与技术脉络及其在通信网络安全中的典型应用。行为智能将AI对抗从静态的特征空间推向动态的策略空间,使得智能体能够理解对手的“所想”与“将做”。在行为智能整体架构中,博弈建模提供了理论框架,行为预测与对手建模赋予系统思考理解能力,而欺骗构成了对抗的博弈战场。未来6G网络将使通信基础设施成为AI对抗的前沿战场,行为智能无疑将重塑防御范式,期待业界建立更系统的评估基准和理论体系,为未来智能网络的自主安全奠定坚实基础。

参考文献:

- [1] Exabeam. Behavior Intelligence for the Agentic Enterprise [EB/OL]. [2026-05-10]. <https://www.exabeam.com/resources/data-sheets/behavior-intelligence-for-the-agentic-enterprise/>.
- [2] Kim M H. Structured Cognitive Loop for Behavioral Intelligence in Large Language Model Agents[J]. arXiv:2510.05107, 2026.
- [3] Park P S, Goldstein S, O'Gara A, et al. AI deception: A survey of examples, risks, and potential solutions [J]. Patterns, 2024, 5 (5): 100988.
- [4] Wu Y, Li T L, Liu H. Behavioral Integrity Verification for AI Agent Skills[J]. arXiv:2605.11770, 2026.
- [5] Jo K, Choi J, Kim K, et al. Artificial Behavior Intelligence: Technology, Challenges, and Future Directions [C]//Proceedings of the 2025 IEEE International Conference on Human System Interaction (HSI 2025). IEEE, 2025: 1-10.
- [6] Zhou J, Chen Y, Shi Y, et al. SocialEval: Evaluating Social Intelli-

gence of Large Language Models [C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025). Vienna, Austria, 2025: 30958-31012.

- [7] Aljović A, Lin Z W, Wang W B, et al. An autonomous AI agent for universal behavior analysis [J]. bioRxiv, 2025.
- [8] Chalkiadakis G, Akasiadis C, Koresis G, et al. Graph Neural Networks, Deep Reinforcement Learning and Probabilistic Topic Modeling for Strategic Multiagent Settings [J]. arXiv:2511.10501, 2026.
- [9] 程恺, 张金鹏, 邵天浩, 等. 智能博弈领域中的对手建模方法综述 [J]. 计算机技术与发展, 2025, 35(9): 1-8.
- [10] Ganzfried S. Consistent Opponent Modeling in Imperfect-Information Games [J]. arXiv:2508.17671, 2025.
- [11] Huang Y, Sun Y, Zhang Y, et al. DeceptionBench: A Comprehensive Benchmark for AI Deception Behaviors in Real-world Scenarios [C]//Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025). 2025.
- [12] Xu Y, Li M, Shen Z, et al. AT-Field: Rethinking the Games in Adversarial Training [C]//Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI 2026). 2026, 40(32): 27368-27376.
- [13] Sun Z, Zhang Z, Liang D, et al. "To Survive, I Must Defect": Jail-breaking LLMs via the Game-Theory Scenarios [J]. arXiv: 2511.16278, 2025.
- [14] Albrecht S V, Stone P. Autonomous agents modelling other agents: A comprehensive survey and open problems [J]. Artificial Intelligence, 2018, 258: 66-95.
- [15] Jeloka B, Guan Y, Tsiotras P. Learning Large-Scale Competitive Team Behaviors with Mean-Field Interactions and Online Opponent Modeling [J]. arXiv:2504.21164, 2026.
- [16] Curvo P M P. The Traitors: Deception and Trust in Multi-Agent Language Model Simulations [J]. arXiv:2505.12923, 2025.
- [17] Hackett W, Garraghan P. Behind the Refusal: Determining Guardrail Activation via Behavioral Monitoring [J]. arXiv:2607.02121, 2026.
- [18] Ferrel V. HBEE: Human Behavioral Entropy Engine—Pre-Registered Multi-Agent LLM Simulation of Peer-Suspicion-Based Detection Inversion [J]. arXiv:2605.07472, 2026.
- [19] Munir M S, Proddatoori S, Muralidhara M, et al. A Trust-By-Learning Framework for Secure 6G Wireless Networks Under Native Generative AI Attacks [J]. IEEE Open Journal of the Communications Society, 2025, 6: 6045-6065.
- [20] Castro S R, Campbell R, Lau N, et al. Large Language Models are Autonomous Cyber Defenders [C]//Proceedings of the IEEE CAI 2025 Adaptive CyberDefense Workshop. IEEE, 2025.
- [21] Next Steps in Cyber Blue Team Automation—Leveraging the Power of LLMs [C]//2025 IEEE/IFIP Network Operations and Management Symposium (NOMS). IEEE, 2025.

作者简介:

陈蓉, 工程师, 硕士, 主要从事网络安全研究与规划工作; 侯捷, 工程师, 硕士, 主要从事网络安全研究与规划工作; 徐雷, 正高级工程师, 博士, 主要从事网络安全管理工作。