全国大模型与决策智能大会

（七）安全可信智能分论坛

论坛概要：本专题将深入探讨大模型的安全与信任问题，旨在通过综合讨论关键技术和策略，构建更加可靠和负责任的大规模人工智能系统。本专题的探讨重点是提升大模型在实际应用中的鲁棒性、可解释性和公平性，以确保其操作的稳定性和安全性，并增强公众对这些先进技术的信任。此外，论坛还将专注于识别和讨论 AI 系统在部署过程中可能遇到的安全漏洞和风险。我们期望本次论坛能够推动安全可信智能技术的研究前沿，促进其在学术和研究领域的实际应用。

1.论坛主席简介

姓名：操晓春

中山大学信息学部副主任、网络空间安全学院院长，国家杰出青年/优秀青年基金获得者。主要从事人工智能基础研究和网络空间内容安全应用研究；发表 ACM/IEEE 汇刊 100 余篇，CCF-A 类期刊及会议长文文章 160 余篇；Google 引用 20000 余次，H-index 65；获得省部级一等奖和二等奖各 1 项。现兼任 TPAMI 的 Associate Editor、TIP 的 Senior Area Editor、电子学报的编委，曾兼任 TMM 和 TCSVT 的的 Associate Editor，10 余次兼 NeurIPS/ICCV/CVPR/IJCAI/ACMMM 的 Area/Track Chairs。指导博士生获得中国电子学会优博、CCF 优博、中科院优博论文 3 篇次；指导的研究生有 4 人入选国家级人才计划。

2.论坛内容

序号	报告人	报告名称	职称/职务	工作单位
1	沈超	AI 大模型的安全与隐私风险	教授/副处长	西安交通大学
2	黄晓霖	深度神经网络中的 随机性与稳健性提升	教授/副主任	上海交通大学 自动化系
3	张卫明	生成式人工智能驱动的 认知安全	教授/副院长	中国科学技术大学
4	张拳石	神经网络是否可以被严谨地解释清楚？以及可解释性技术在大模型上的应用落地	副教授	上海交通大学电子信息与电气工程学院
5	梁思源	多模态基础模型中后门攻防的 全链路研究	博士后	新加坡国立大学

3.报告人简介

姓名：沈超

报告题目：AI 大模型的安全与隐私风险

西安交通大学人才办副处长、二级教授，教育部长江学者特聘教授，教育部创新团队负责人，国家重点研发计划首席科学家，国防基础加强计划首席科学家，重点研发计划“先进计算与新兴软件”重点专项指南专家组成员。主要从事智能系统可信、安全、控制与测试的研究工作，发表学术刊物 180 余篇，获最佳论文奖 9 次。牵头获陕西省科学技术一等奖、中国自动化学会科学技术一等奖、达摩院青橙奖、霍英东教师一等奖、MIT TR35 China、国家优秀青年科学基金、IEEE SMC Early Career Award、陕西省五四青年奖章等。主持国家重大、重点、国际(地区)合作等项目 30 余项，制定国内外标准 5 项，多份建言被中央办公厅等采纳。担任 IEEE TDSC、TCYB 汇刊等 10 余个国际期刊编委、IEEE Xi'an SMC&CS 主席、ACM SIGSAC China 副主席、中国人工智能学会组织工委副主任等。

姓名：黄晓霖

报告题目：深度神经网络中的随机性与稳健性提升

上海交通大学教授，博士生导师，青年千人计划入选者。分别在西安交通大学、清华大学获得工学学士、工学博士学位。面向深度学习泛化性，黄晓霖在函数空间、优化方法、对抗攻击等方面进行了持续研究，发表机器学习领域顶刊JMLR论文5篇;IEEE TPAMI论文8篇,在Nature Review Methods Premiers 发表综述 1 篇，并多次在领域重要会议如 NeurIPS, ICLR, CVPR, MICCAI 等做学术汇报。目前担任 Machine Learning 的 Action Editor、ICCV 的 Area Chair、AAAI 的 Senior TPC member 等；承担科技部重点研发课题、自然基金面上项目、上海市科委人工智能专项等，并与华为、美敦力等公司进行了长期的合作。

姓名：张卫明

报告题目：生成式人工智能驱动的认知安全

中国科学技术大学教授、博导，网络空间安全学院副院长。主要研究兴趣包括信息隐藏和人工智能安全。已在国际著名学术期刊和会议 IEEE TIT、TPAMI、TIFS、TIP、CVPR、S&P、ICCV、NeurIPS、AAAI 等发表论文 200 多篇。主持基础加强重点项目、国家自然科学基金重点、国家重点研发课题、国家 863 等项目 20 余项。获得军队科技进步一等奖、安徽省自然科学奖一等奖、安徽省教学成果特等奖、ACM SIGSOFT 杰出论文奖。入选 2021 年长三角人工智能十大杰出人物。

姓名：张拳石

报告题目：神经网络是否可以被严谨地解释清楚？以及可解释性技术在大模型上的应用落地

上海交通大学电院计算机科学与工程系，长聘教轨副教授，博士生导师，入选国家级海外高层次人才引进计划，获 ACM China 新星奖。他于 2014 年获得日本东京大学博士学位，于 2014-2018 年在加州大学洛杉矶分校（UCLA）从事博士后研究。张拳石在神经网络可解释性方向取得了多项具有国际影响力的创新性成果。张拳石承担了 TMLR 的 Action Editor，CCF-A 类会议 IJCAI 2020 和 IJCAI 2021 的可解释性方向的 Tutorial，并先后担任了 AAAI 2019, CVPR 2019, ICML 2021 大会可解释性方向的分论坛主席。

姓名：梁思源

报告题目：多模态基础模型中后门攻防的全链路研究

新加坡国立大学计算机学院 Research Fellow，入选腾讯犀牛鸟精英人才计划，获中国科学院院长优秀奖、国家奖学金、DAAD AInet 奖学金。她于 2023 年获得中国科学院大学网络空间安全学院博士学位，主要研究兴趣包括计算机视觉和人工智能鲁棒性。已在国际知名学术期刊和会议，如 IEEE TIFS、CVPR、USENIX Security、ICCV、ECCV、AAAI、ACM MM、ICLR 等，发表 20 余篇论文，多篇论文被选为顶级会议的 Oral、Highlight、Spotlight 报告并获得了 CVPR 2024 关于基础模型鲁棒性专题研讨会的最佳论文奖。此外，她曾三次在 A 类会议上担任“智能安全”系列主题国际研讨会及竞赛的论坛主席。