

9:00 – 9:40, Tuesday, Aug. 18, 2026

Plenary Talk

Approaches to Safe and Trustworthy Physical Agents

Jianru Xue

Professor, Xi'an Jiaotong University



Biography: Jianru Xue, PhD, Professor. He received his PhD degrees from Xi'an Jiaotong University in 2003. He has joined Xi'an Jiaotong University since 1999. He had worked in FujiXerox, Tokyo, Japan, from 2002 to 2003. In 2008-2009, he had visited University of California, Los Angeles. His research interests include pattern recognition and machine learning, autonomous systems and hybrid augmented intelligence. He and his team won National Natural Science Award (second class) in 2016, National Technology Invention Award in 2007, the IEEE ITSS Institute Lead Award in 2014. He has published 100+ papers in top cited journals and conferences including IEEE TPAMI/TIP/TSMCB, CVPR, ICCV, ECCV, ACM MM, RSS, ICRA, IROS etc.

Abstract: Physical agents face an urgent need to collaborate efficiently with humans in open, dynamic environments. To address the uncertainties inherent in these settings, researchers propose designing policies by integrating Deep Reinforcement Learning (DRL) with Large Language Models (LLMs). The continuous emergence of physical agents with diverse morphologies is generating global, diverse, and large-scale closed-loop data, significantly propelling the development of data-driven autonomous policy learning. However, concerns regarding the safety and trustworthiness of autonomous behaviors are just beginning to surface—the tip of the iceberg. This report focuses on approaches for designing safe and trustworthy autonomous physical agents under uncertainty from two perspectives: endogenous safety and safety filters. We propose combining traditional controller design frameworks grounded in mathematical models with the emerging paradigm of LLM-integrated DRL. The report highlights our team's recent progress in these areas and outlines several fundamental research questions worthy of further consideration.