上海交通大学自动化系
Department of Automation
Shanghai Jiao Tong University

系统控制与信息处理教育部重点实验室
Key Laboratory of System Control and Information Processing, Ministry of Education

学术报告会

时间：**2023年9月6日 10:00**
地点：电信群楼**2-410**会议室

Towards Trustworthy AI: Sandboxing AI-based Controllers in Cyber-Physical Systems

Bingzhuo Zhong

加州大学伯克利分校，博士后

Abstract:

The past decades have witnessed remarkable achievements in artificial intelligence (AI) in many domains, such as natural language processing and image recognition. In the near future, many AI-based controllers are also expected to be deployed in modern Cyber-Physical System (CPSs) to accomplish complex control missions; typical scenarios include autonomous vehicles and smart buildings. However, the verification of many AI-based controllers, particularly those developed based on deep neural networks, is a challenging task. Meanwhile, modern CPSs are typically safety-critical and prone to various security threats due to the tight interaction and frequent information exchange between their cyber and physical components. Therefore, the difficulties in verifying AI-based controllers may lead to disastrous consequences for real-life CPSs in terms of safety and security concerns. In this talk, I will introduce how to design a system-level, secure-by-construction architecture to sandbox AI-based controllers leveraging recent results in hybrid systems. By sandboxing AI-based controllers with the proposed architecture, safety and security guarantees can be provided for CPSs, while formal verification over those AI-based controllers deployed in the CPSs is not necessary.

Biography:

Bingzhuo Zhong is currently a postdoctoral researcher at the University of California, Berkeley, and University of Colorado Boulder, USA. He received a doctoral degree (Dr. rer. nat.) in computer science (summa cum laude) in 2023 from the Technical University of Munich, Germany. He received an M.Sc. degree in mechanical engineering in 2018 from the Technical University of Munich, Germany, and a B.Eng.degree in vehicle engineering in 2016 from Tongji University, China. His research interests include trustworthy artificial intelligence, verification and synthesis of hybrid system, safety and security of Cyber-Physical Systems, data-driven control, and digital twin in smart control engineering.