Speaker:
Professor Yorie Nakahira (Carnegie Mellon University)
https://www.ece.cmu.edu/directory/bios/nakahira-yorie.html

Date and Place:
6th, June, 2023, 15:00

Faculty of Engineering Integrated Research Bldg, Room 213, Kyoto University
(京都大学工学部総合校舎 213）

Title:
Safety of intelligent systems operating in uncertain and interactive environments

Abstract:

Autonomous systems utilizing learning-based techniques must safely operate in uncertain and interactive environments. In this talk, we will overview our recent projects focusing on the safety of such systems.

One critical aspect of safe decision-making is to account for strategic human behaviors. While safe control frameworks are often designed to behave safely even in worst-case human uncertainties, this can encourage humans to behave more aggressively and result in greater risk for everyone. Here, we will present a framework to formally investigate situations in which conservative safe control methods compromise safety. Our analysis suggests a need to rethink the safe control method in highly interactive situations.

Another critical aspect is to deal with latent risks and extreme situations, e.g., driving on a slippery and occluded mountain road. Here, we will introduce a probabilistic safety certificate that can certify a data-driven controller while accounting for latent risks and uncertainty in changing system dynamics. This certificate integrates reachability- and invariance-based approaches via a new notion of forward invariance to exploit the advantages of both approaches. The proposed method has three features. First, it systematically mediates behaviors based on the levels of latent risks and interaction models. Second, it guarantees long-term safety using a computationally efficient (myopic) controller. Third, it allows multiple agents to collaboratively ensure system-wide safety specifications even if each agent does not have sufficient information to evaluate the specifications.

Finally, many safe control and learning techniques, including the ones introduced above, require an accurate estimation of long-term risk probabilities and their gradients. However, the evaluation of long-term risks and rare events requires heavy computation. Here, we will show that a physics-informed neural network can be used to estimate long-term risk probabilities with provable guarantees. The neural network can be trained using samples from short-term trajectories and partial states to accurately estimate longer-term risk probabilities of unseen states, and do so with provable accuracy under mild assumptions.